
THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

# A Semantic e-Science Platform for 20th Century Paint Conservation

Suleiman Abdullah Suleiman Odat

*A thesis submitted for the degree of Doctor of Philosophy at*

*The University of Queensland in September 2014*

School of Information Technology and Electrical Engineering

**Abstract**

Throughout the twentieth century, artists in Australia and across the Southeast Asia-Pacific region have enthusiastically embraced new materials (synthetic media, new pigments, dyes and additives). But compared to traditional artists' paints, these new materials have affected paint handling and paint stability. These new materials have also resulted in a lack of understanding of the preservation issues associated with the resulting artworks. As a result, today's collectors, curators and conservators are confronted with significant material-based preservation questions associated with 20th century art/paint preservation – but they lack the sustained and integrated knowledge-base to inform their decision making.

In order to understand the causes of paint degradation and the best preservation and treatment approaches, conservators need access to a wide range of distributed and cross-disciplinary datasets. They need to access: historical and provenance data associated with individual paintings; information about artistic techniques; paint chemistry databases; publications on preservation treatments and previous research; and collaborative, but secure Web-based tools for capturing, sharing and discussing condition reports, deterioration mechanisms, and characterisation/imaging data (e.g., Scanning Electron Microscopy, Transmission Electron Microscopy, Fourier Transform Infrared Spectroscopy, and X-Ray Diffraction).

The aim of this research project is to develop and apply the latest information integration, data management and Semantic Web technologies to build an effective, scalable, extensible, flexible and portable knowledge-base for 20th century art/paint preservation using an approach that enhances the discoverability and re-use of knowledge. The aims of this project are to develop an e-Research platform for art conservators by tackling the following steps/objectives:

- Develop an Ontology of Paintings and PReservation of Art (OPPRA) that will link and integrate terms from standard and disciplinary ontologies (e.g., CIDOC-CRM, OreChem and OAI-ORE) with existing, relevant thesauri (e.g., Getty Art and Architecture Thesaurus and CAMEO: Conservation and Art Material Encyclopedia Online) and new ontologies (e.g., describing types of paint deterioration);

- Use a number of case studies to evaluate OPPRA's ability to capture the detailed workflows and outputs associated with paint conservation experiments (e.g., sampling method, experimental processes and characterisation data);
- Apply and optimise a combination of semantic tagging and machine learning approaches to extract structured knowledge (compliant with OPPRA) from free-text publications on paint conservation – so it can be shared, integrated, compared and re-used;
- Evaluate OPPRA's ability to integrate experimental datasets, structured knowledge extracted from free-text publications and external public relevant databases (e.g., on paint chemistry), to answer a set of advanced, exemplar (SPARQL) queries specified by art conservators;
- Evaluate OWL-DL for inferencing and extracting new facts from the integrated knowledge base (generated from integrating experimental data capture, structured knowledge extracted from past publications and public relevant databases) in order to answer advanced queries specified by art conservators.

**Declaration by author**

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the General Award Rules of The University of Queensland, immediately made available for research and study in accordance with the *Copyright Act 1968*.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

## Publications during candidature

### Refereed Journal Articles

Odat, S., Groza, T. & Hunter, J. 2014. **Extracting Structured Data from Publications in the Art Conservation Domain.** Literary and Linguistic Computing, Oxford Journals. doi: 10.1093/llc/fqu002. – Mostly incorporated as Chapter 7.

### Peer-Reviewed Conference Papers

Odat, S. & Hunter, J. 2012. **Developing an online collaborative knowledge-base for art conservators**. In: Proceedings of the 3rd APTCCARN Meeting in 2012. *3rd APTCCARN Meeting in 2012*, Thailand. 23-25 April 2012. – Partially incorporated as paragraphs in Chapters 1, 4-5 and 8.

Hunter, J. & Odat, S. 2011. **Building a semantic knowledge base for painting conservators**. In: *7th IEEE International Conference on e-Science*, Stockholm, Sweden. 5-8 December 2011.– Partially incorporated as paragraphs in Chapters 1-4 and 6.

### Conference Papers (Paper Extracts)

Odat, S. & Hunter, J. 2013. **Providing Decision Support Tools and a Shared Online Knowledge Base for Paint Conservators**. In: *1st International Materials Informatics Conference*, Melbourne, Australia. 31 January – 06 February 2013. – Partially incorporated in paragraphs throughout the thesis.

Odat, S. & Hunter, J. 2012. **Providing Decision Support Tools for Art Conservators through a Shared Knowledge Base**. In: The Meaning of Materials in Modern and Contemporary Art. *13th AICCM Paintings Special Interest Group Symposium*, Brisbane, Australia. 10-11 December 2012. – Partially incorporated in paragraphs throughout the thesis.

Hunter, J. & Odat, S. 2012. **Building an Online Collaborative Knowledge-Base for Art Conservators**. In: *23rd International CODATA Conference*, Taipei Taiwan. 28-31 October 2012. – Partially incorporated in paragraphs throughout the thesis.

Hunter, J., Odat, S., Osmond, G. & Drennan, J.2012. **Building a semantic knowledge-base for painting conservators in Asia-Pacific**. In: *Digital Humanities Australasia (DHA)*, Canberra Australia. 28-30 March 2012. – Partially incorporated in paragraphs throughout the thesis.

Odat, S. & Hunter, J. 2011. **A Semantic e-Science Platform for 20th Century Paint Conservation**. In: *Scoping the future of cultural enrichment through cultural materials conservation*, Melbourne, Australia. 16 June 2011. – Partially incorporated in paragraphs throughout the thesis.

Odat, S. & Hunter, J. 2010. **Data management and integration services for 20th century paint conservation**. In: Dialogues with Artists. *12th AICCM Paintings Special Interest Group Symposium*, Adelaide, Australia. 21-22 October 2010. – Partially incorporated in paragraphs throughout the thesis.

Hunter, J. & Odat, S. 2010. **eResearch services for 20th century paint conservation**. In: *Universitas 21 Digital Humanities Conference*, University of Birmingham, UK. 15-17 September 2010. – Partially incorporated in paragraphs throughout the thesis.

**Posters**

Odat, S. & Hunter, J. **Knowledge Management Services for an Online Community of Painting Conservators**. eResearch Australasia, Sydney Masonic Centre, Sydney, Australia. 28-31 October 2012.

Odat, S. & Hunter, J. **Knowledge Management Services for an Online Community of Painting Conservators**. Innovation Expo, The University of Queensland, Brisbane, Australia. 30 October 2012.

Odat, S. & Hunter, J. **An Online Knowledge-base for 20th Century Painting Conservation**. UQ Day for Digital Humanities, The University of Queensland, Brisbane, Australia. 24 September 2010.

## Publications included in this thesis

Odat, S., Groza, T. & Hunter, J. 2013. **Extracting Structured Data from Publications in the Art Conservation Domain.** Literary and Linguistic Computing, Oxford Journals.doi: 10.1093/llc/fqu002.– Mostly incorporated as Chapter 7.

| Contributor | Statement of contribution |
|---|---|
| Author Odat, S. (Candidate) | Implemented Back/Front-end (100%)<br><br>Designed experiments (100%)<br><br>Wrote and edited paper (65%) |
| Author Groza, T. | Wrote and edited paper (10%) |
| Author Hunter, J. | Wrote and edited paper (25%) |

Odat, S. & Hunter, J. 2012. **Developing an online collaborative knowledge-base for art conservators**. In: Proceedings of the 3rd APTCCARN Meeting in 2012. *3rd APTCCARN Meeting in 2012*, Thailand. 23-25 April 2012. – Partially incorporated as paragraphs in Chapters 1, 4-5 and 8.

| Contributor | Statement of contribution |
|---|---|
| Author Odat, S. (Candidate) | Implemented Back/Front-end (100%)<br><br>Wrote and edited paper (50%) |
| Author Hunter, J. | Wrote and edited paper (50%) |

Hunter, J. & Odat, S. 2011. **Building a semantic knowledge base for painting conservators**. In: *7th IEEE International Conference on e-Science*, Stockholm, Sweden. 5-8 December 2011.– Partially incorporated as paragraphs in Chapters 1-4 and 6.

| Contributor | Statement of contribution |
|---|---|
| Author Hunter, J. | Wrote and edited paper (55%) |
| Author Odat, S. (Candidate) | Implemented Back/Front-end (100%)<br><br>Wrote and edited paper (45%) |

Odat, S. & Hunter, J. 2012.**Developing an online collaborative knowledge-base for art conservators**. In: Proceedings of the 3rd APTCCARN Meeting in 2012. *3rd APTCCARN Meeting in 2012*, Thailand. 23-25 April 2012. – Partially incorporated as paragraphs in Chapters 1, 4-5 and 8.

| Contributor | Statement of contribution |
|---|---|
| Author Odat, S. (Candidate) | Implemented Back/Front-end (100%)<br><br>Wrote and edited paper (50%) |
| Author Hunter, J. | Wrote and edited paper (50%) |

Odat, S. & Hunter, J. 2013. **Providing Decision Support Tools and a Shared Online Knowledge Base for Paint Conservators**. In: *1st International Materials Informatics Conference*, Melbourne, Australia. 31 January – 06 February 2013. – Partially incorporated in paragraphs throughout the thesis.

| Contributor | Statement of contribution |
|---|---|
| Author Odat, S. (Candidate) | Wrote and edited paper (50%) |
| Author Hunter, J. | Wrote and edited paper (50%) |

Odat, S. & Hunter, J. 2012. **Providing Decision Support Tools for Art Conservators through a Shared Knowledge Base**. In: The Meaning of Materials in Modern and Contemporary Art. *13th AICCM Paintings Special Interest Group Symposium*, Brisbane, Australia. 10-11 December 2012. – Partially incorporated in paragraphs throughout the thesis.

| Contributor | Statement of contribution |
|---|---|
| Author Odat, S. (Candidate) | Wrote and edited paper (50%) |
| Author Hunter, J. | Wrote and edited paper (50%) |

Hunter, J. & Odat, S. 2012. **Building an Online Collaborative Knowledge-Base for Art Conservators**. In: *23rd International CODATA Conference*, Taipei Taiwan. 28-31 October 2012. – Partially incorporated in paragraphs throughout the thesis.

| Contributor | Statement of contribution |
|---|---|
| Author Hunter, J. | Wrote and edited paper (50%) |
| Author Odat, S. (Candidate) | Wrote and edited paper (50%) |

Hunter, J., Odat, S., Osmond, G. & Drennan, J. 2012.**Building a semantic knowledge-base for painting conservators in Asia-Pacific**. In: *Digital Humanities Australasia (DHA)*, Canberra Australia. 28-30 March 2012. – Partially incorporated in paragraphs throughout the thesis.

| Contributor | Statement of contribution |
|---|---|
| Author Hunter, J. | Wrote and edited paper (45%) |
| Author Odat, S. (Candidate) | Wrote and edited paper (45%) |
| Author Osmond, G. | Minor edits (5%) |
| Author Drennan,, J. | Minor edits (5%) |

Odat, S. & Hunter, J. 2011. **A Semantic e-Science Platform for 20th Century Paint Conservation**. In: *Scoping the future of cultural enrichment through cultural materials conservation*, Melbourne, Australia. 16 June 2011. – Partially incorporated in paragraphs throughout the thesis.

| Contributor | Statement of contribution |
|---|---|
| Author Odat, S. (Candidate) | Wrote and edited paper (50%) |
| Author Hunter, J. | Wrote and edited paper (50%) |

Odat, S. & Hunter, J. 2010. **Data management and integration services for 20th century paint conservation**. In: Dialogues with Artists. *12th AICCM Paintings Special Interest Group Symposium*, Adelaide, Australia. 21-22 October 2010. – Partially incorporated in paragraphs throughout the thesis.

| Contributor | Statement of contribution |
|---|---|
| Author Odat, S. (Candidate) | Wrote and edited paper (50%) |
| Author Hunter, J. | Wrote and edited paper (50%) |

Hunter, J. & Odat, S. 2010. **eResearch services for 20th century paint conservation**. In: *Universitas 21 Digital Humanities Conference*, University of Birmingham, UK. 15-17 September 2010. – Partially incorporated in paragraphs throughout the thesis.

| Contributor | Statement of contribution |
|---|---|
| Author Hunter, J. | Wrote and edited paper (50%) |
| Author Odat, S. (Candidate) | Wrote and edited paper (50%) |

**Contributions by others to the thesis**

No contributions by others

**Statement of parts of the thesis submitted to qualify for the award of another degree**

None

**Acknowledgements**

This thesis would not have been possible without the encouragement and support of many people.

Firstly, I owe my deepest gratitude to my supervisor, Professor Jane Hunter. She has guided, encouraged, and worked with me throughout my thesis with her patience and knowledge. From her, I have learned very much professionally and personally. I would also like to thank Professor John Drennan for the effort he put into guiding me through the process. They have become not only my PhD advisors, but also my mentors and role models for my future career.

Secondly, I am grateful to all the people I collaborated with in the 20th Century in Paint project for the valuable discussions and all the things I learned from them. They have provided me with a unique opportunity to conduct research in an area I have always been interested in. I am grateful to Gillian Osmond, Paula Dredge, Associate Professor Robyn Sloggett, and Dr. Nicole Tse for their support, case studies and corpus selection. Funding for this work comes from the Australian Research Council (ARC) Linkage Program for the 20th Century in Paint (Project LP0883309) to whom I am grateful for their support.

Thirdly, thank you to eResearch Lab co-workers (School of Information Technology and Electrical Engineering, The University of Queensland) who provided a challenging, interesting and entertaining environment to work in. The feedback provided by many of them at various stages throughout this work was much appreciated. Particular thanks to Dr. Tudor Groza who gave enthusiastic advice and support for the text analysis approach in Chapter 7, and throughout the structure of this thesis. I must also acknowledge Anna Gerber, Carol Owen, Charles Brooking, Dr Stephen Crawley, Dr Nigel Ward, Dr Peter Ansell and Damien Ayers for their invaluable constructive feedback and fruitful discussions. Furthermore, I would like to thank my colleagues and friends Chih-Hao (David) Yu, Lianli (Juana) Gao, Razan Paul, Hasti Ziaimatin, Kutila Gunasekera and Hamed Hassanzadeh, with whom I had a wonderful time here at the eResearch Lab.

Fourthly, I want to thank my family to whom I owe everything. Thanks to Dad for being the first and the greatest teacher in my life, thanks to Mum for dedicating half

of your life raising me. And many thanks to my three brothers and six sisters for their encouragement and support during my study. I dedicate this achievement to my family.

Last, but foremost, the experience of parenting my dearest daughter Nadia, over the past 5 years has given meaning to my life that cannot be matched by my career achievements. Thank you Nadia for being a constant source of joy, pride, and inspiration, I couldn't have done it without you being here with me. Saving the most important acknowledgment for the end, I owe so much to my soul mate, best friend, intellectual inspiration, pillar of strength, most valuable critic, and wife Faith. As you have heard me say over and over again, I lead a blessed life, and the completion of this thesis is further proof of that. But the most blessed event in my life is the time I met you and the continual blessing of sharing my life with you fills me with constant joy.

**Keywords**

semantic web, material informatics, owl, provenance, paint conservation

**Australian and New Zealand Standard Research Classifications (ANZSRC)**

ANZSRC code: 080608, Information Systems Development Methodologies, 100%

**Fields of Research (FoR) Classification**

FoR code: 0899, Other Information and Computing Sciences, 100%

## Table of Contents

## List of Figures

## List of Tables

## List of Abbreviations

AAT: Art and Architecture Thesaurus (Getty Vocabularies)
AICCM: Australian Institute for the Conservation of Cultural Material
AMA: Archive Mapper for Archaeology
API: Application Programming Interface
APTCCARN: Asia Pacific Twentieth Century Conservation Art Research Network
ATR: Attenuated Total Reflectance
CAMEO: Conservation and Art Material Encyclopedia Online
CIDOC-CRM: Conceptual Reference Model (developed by the International Council of Museums' International Committee for Documentation)
DAAO: Dictionary of Australian Artists Online
DALI: Data Aggregation and Linking Interface
EDX: Energy-Dispersive X-Ray
FTIR: Fourier Transform Infrared spectroscopy
Gc: Gas Chromatography
INCCA: International Network for the Conservation of Contemporary Art
IRUG: Infrared and Raman Users Group - Spectral Database
JAIC: Journal of the American Institute of Conservation
ML: Machine Learning
Ms: Mass Spectrometry
NE: Named Entity / NER: Named Entity Recognition
NIST: National Institute of Standards and Technology (Chemistry WebBook)
OAI-ORE: Open Archive Initiative – Object Reuse and Exchange
OPPRA: Ontology of Paintings and PReservation of Art
OWL: Web Ontology Language
PoS: Part-of-Speech
Py: Pyrolysis Gas
RDF: Resource Description Framework
RE: Relationship Extraction
SEM: Scanning Electron Microscopy
SVM: Support Vector Machine
TEM: Transmission Electron Microscopy
URI: Uniform Resource Identifier
UV: Ultraviolet
W&N: Winsor & Newton 19th Century Archive
XACML: eXtensible Access Control Markup Language
XML: eXtensible Markup Language
XRD: X-Ray Diffraction
XRF: X-Ray Fluorescence

# Chapter 1

# Introduction

## 1.1. Background

The aim of the Asia Pacific Twentieth Century Conservation Art Research Network (APTCCARN, 2010) is to explore the preservation of twentieth century paintings in Asia and the Pacific. Modern paintings, in particular, are highly susceptible to problems such as aging, cracking and fading due to the increased instability of modern synthetic organic pigments and paint formulations. Therefore, it is vitally important that these modern pigments, along with their synthetic binders and additives, are characterised before and after problems arise in order to determine the optimum conservation treatments and environmental conditions for their storage, display and transport. Non-invasive analytical methods, such as Scanning Electron Microscopy – SEM, Energy-Dispersive X-ray – EDX, X-Ray Diffraction – XRD and Raman spectroscopy, facilitate the improved identification of modern synthetic organic pigments in acrylic and alkyd paint formulations and oil media, as well as improved understanding of the reactions that they may undergo over time or with exposure to humidity, light and high temperatures.

Due to the increased availability of such sophisticated techniques, painting conservation has evolved into a highly multi-disciplinary research topic that requires the integration of data, information and knowledge about a number of areas including: art history (artworks, artists and artistic techniques); the physical and chemical properties of paint and pigments; and paint conservation techniques (the cleaning, preservation and characterisation methods) that can be used to determine the precise cause of the degradation or discoloration that is occurring and the

optimum treatment to remove or limit the effects. The art conservation field has adopted sophisticated characterisation and imaging methods in the digital age, and the result of this is the need for new Information and Communication Technologies to store, curate, integrate, analyse, visualise and query large volumes of heterogeneous, distributed data and information.

The high-level objective of the research presented in this thesis is to work with the paint conservators and materials scientists involved in the APTCCARN network and the 20th Century in Paint project (20thcpaint, 2010b) – to investigate optimum information integration and analysis technologies to enable an online network of art conservators, curators and materials scientists in the Asia-Pacific region to advance their understanding of the conservation of 20th century paintings and to exchange information on paintings, paint materials, suppliers, artists and art conservation techniques, using agreed standards and Semantic Web approaches (Crofts et al., 2010, Lagoze, 2009, Lagoze et al., 2008).

### 1.1.1. Requirements for Art Conservation

A number of organizations and networks have previously identified the need for art conservators to adopt improved information/data management and e-Research methods for investigating, documenting, sharing and publishing art conservation research, techniques and discoveries. For example, the Andrew W. Mellon Foundation recently funded the ConservationSpace project (Mellon, 2009) which aims “to develop a shared solution to the problem of documentation management” for the conservation community. ConservationSpace was still in the building phase at the time of this thesis but it is focussed on developing a functional system to support real-world conservators in their day-to-day operations. The scope of this thesis is to investigate and evaluate optimum approaches to support the capture, sharing and dissemination of paint conservation research activities and outcomes.

Within this section, the conservation communities’ key requirements (associated with the capture, storage, interpretation, analysis, sharing, publishing and dissemination of art conservation experimental research and knowledge) are outlined. These specific requirements were identified during a series of workshops that were held by the APTCCARN and 20th Century in Paint project members between 2010 and 2011.

Firstly, art conservators and scientists (working on art conservation) need online repositories where they can store and describe each investigation (e.g., the source of the paint samples, the experimental conditions and the characterisation/analysis results). To maximize discovery, interoperability and re-use, such repositories should use standardised and machine-processable metadata schemas, vocabularies and formats.

Secondly, they need to be able to search, correlate and integrate relevant existing data and information on art materials, paints, paint deterioration mechanisms, paint characterisation data, conservation techniques, provenance and artistic practices. Although a large amount of this information is in private databases and not accessible, the focus of this thesis is on improving access to the significant amount of relevant data that is available through public/online databases, Web sites and related publications.

Thirdly, persistent online identifiers (URLs) are required to ensure long-term access to and unique identification of the associated resources via the Semantic Web – i.e., the artworks, samples, instruments, images, experiments, characterisation results, and publications. Currently the relevant resources are scattered on the Web, and their discoverability and re-use through URLs is unreliable because many resources have not been assigned persistent unique identifiers or the assigned URLs may have been moved, removed or renamed.

Fourthly, conservators and scientists need to be able to protect their results through authenticated access control mechanisms until they are ready to share them with colleagues or publish them. This requirement arises due to the security, privacy and intellectual property restrictions imposed by authors, organisations and collaborating individuals.

Finally, art conservators require integrated and event-aware informatics framework/knowledge-bases that can:

- Enable *social semantic networks* by linking teams of art conservators with information resources that capture/describe: paintings, artistic techniques, art

provenance, samples, experiments, characterisations and preservation treatments;

- Provide more accurate answers to more sophisticated queries than traditional databases such as: *What is the best way to treat zinc oxides occurring in paintings by Rover Thomas? What are the factors that cause or accelerate the occurrence of lead soaps in paintings by R. Godfrey Rivers? What is the best solvent for removing varnish from acrylic paintings that exhibit cracking? List all oil paintings that show cracking due to metal soap formation. Retrieve publications that report the presence of lead soap aggregates in artworks painted using Ripolin (Picasso, Sidney Nolan)*;
- Enable them to compare research outcomes with similar research described in related publications (e.g., a preservation technique applied to an 18$^{th}$ century painting in the Rijksmuseum, Amsterdam with the preservation of a 20$^{th}$ century indigenous painting at the Gallery of Modern Art, Brisbane).

### 1.1.2. Example

In an investigation into the appearance of metal soap formation (surface lump aggregation) in some 19$^{th}$ and early 20$^{th}$ century British and Australian paintings (Osmond et al., 2005), paint samples were characterised using SEM-EDX and UV fluorescence. This characterisation (illustrated in Figure 1.1) showed that zinc was consistently found at the centre of the fluorescent regions, indicating that a white pigment (commonly known as zinc white) contained zinc oxide and was reacting to form organic soaps. These soap compounds have a larger surface area and volume which cracks the original paint.

To better understand these compounds, their long-term stability and how to prevent or reduce their formation, a series of experiments on zinc oxide was conducted to simulate its aging and degradation processes within paints. The workflow of the process involved: mixing samples of zinc oxide with acrylic paints; exposing them to controlled environmental conditions such as UV light, temperature and humidity; analysing the structure and composition of the output, using SEM, Transmission Electron Microscopes – TEM, UV fluorescence and Fourier Transform Infrared

spectroscopy – FTIR; and identifying the presence, nature and extent of zinc soaps – $Zn(C_{18}H_{35}O_2)_2$.

**Figure 1.1: Left – *Woolshed* (New South Wales 1890 – R. Godfrey Rivers – Oil on canvas – 92 x 112cm – Gift of the artist 1895 – Queensland Art Gallery), Centre – Surface macro images showing flaking, bubbling and lead soap formation in selected regions of the painting (child's face), Right – SEM-EDX and UV fluorescence images showing a reaction of zinc white and zinc oxide forming organic soaps**

In order to store, analyse and interpret the results of the experiments, share the results with collaborators and eventually publish the results (both the data together with the traditional textual publication), the art conservator requires an online repository where she can describe each investigation, the source of the paint samples, the experimental conditions and the characterisation results – using standardised, machine-processable metadata schemas, vocabularies and formats. She needs to be able to compare her research outcomes with similar research described in related publications. She also needs to be able to protect her results (through authenticated access control mechanisms) until she is ready to publish them. Finally, she needs to be able to publish unique persistent URLs within her publication that enable readers to retrieve the raw images or spectrographic data.

To assist with these requirements, an integrated and event-aware informatics framework for art/paint preservation (based on a formal machine-processable data model/ontology) is needed. Given this common model, tools are needed to:

- Capture new information in a form that complies with the ontology;
- Extract structured knowledge (based on the ontology) from raw data and text gathered from free-text publications and discussions;

- Allow heterogeneous information sources to be searched, aggregated and analysed, using terms in the ontology;
- Enable semantic inferencing across the harvested knowledge.

## 1.2. Motivation

Several factors motivated this research project within the domain of paintings and art conservation. Firstly, the amount and complexity of information of different types that needs to be stored, accessed, validated, manipulated, managed and used for decision-making is staggering. A tremendous amount of information in the form of raw data is generated from simple archiving, assessments and condition reports, artists' choices of oil paints in the 20$^{th}$ century, the perception of paintings which have aged over time, conservation issues of sensitive painted surfaces (e.g., water, temperature and humidity), experimental data and complex models on the physical properties of oil paints. More specifically, the following information needs to be captured to satisfy the typical information integration and analysis needs of art conservators:

- **Paintings** – title, artist, period, technique, genre, condition, owner, custodian, provenance;
- **Paint** – manufacturer, supplier, year, paint name, identifier, bottle label, type, chemical property (e.g., composition, concentration, acidity and solubility), physical property (e.g., dryness, hardness and resistivity), pigment, additive (e.g., thickener, stabiliser, preservative, surfactant, coalescing solvent and defoamer);
- **Paint decomposition** – type (e.g., cracking, peeling, fading, discoloration and mould growth), cause (e.g., humidity, light, temperature, water, artistic technique) and physical/chemical process/reaction;
- **Paint analysis method** – SEM, TEM, FTIR, Raman, XRD, XRF, EDX, Pyrolysis Gas Chromatography Mass Spectrometry – Py-Gc-MS and Synchrotron radiation;
- **Paint conservation/preservation treatment** – cleaning, protective coating, environmental conditions;
- **Experiment** – experimenter, objective, paint sample, parameter, result and data (document, observation, hypothesis, finding, etc).

Secondly, although it is possible to find some concentrated authoritative collections of information on this topic on the Web (e.g., Journal of the American Institute of Conservation – JAIC, Smithsonian Museum Conservation Institute, Getty Conservation and Research Institutes, Conservation and Art Material Encyclopedia Online – CAMEO, and Forbes Pigment database), the relevant information is however difficult to extract, re-use, interpret, correlate or compare because it is:

- Highly heterogeneous – for example, organisational/disciplinary approaches (art conservation, materials chemistry and information and characterisation science);
- Embedded within disparate databases – for example, collections, artists, materials, chemicals and spectra;
- Hidden within highly unstructured textual documents – for example, publications, discussions and technical reports;
- Expressed using different:
    - Terminologies – for example, measurement units, synonyms and chemical identifiers/structures;
    - Data formats – for example, 2D (manuscripts, paintings and photos), 3D (digital objects), video (interviews, exhibitions, performances and artistic techniques), audio (songs, stories, oral history) and virtual reality (animated walkthroughs and advanced computer graphics);
    - Security, privacy, confidentiality and intellectual property agreements – for example, provenance, condition reports and mistakes kept hidden (conservators may be reluctant to admit to mistakes or to share case studies that document errors and help to prevent future similar mistakes being made).

Finally, previous approaches to the construction of semantic knowledge and documenting the physical (e.g., movement, exhibition, condition assessment and treatment) and digital (e.g., sampling, experiment, characterisation and results) provenance of artworks lack the standardised models, ontologies, frameworks, terminologies and machine-processable descriptions of preservation methods (Green and Mustalish, 2009, Hohmann and Schiemann, 2013, Krafft et al., 2010, Pirró et al., 2010, Schmidt et al., 2011).

To overcome these challenges, the development of an Ontology of Paintings and PReservation of Art – OPPRA is proposed in this research. The ontology is a formal and explicit model that enables the:

- Integration of relevant knowledge sources that are distributed across the Web;
- Documentation of experiments that are investigating paint degradation;
- Documentation of the physical and digital provenance of paintings;
- Storage of structured knowledge that is extracted from publications;
- Linking experimental data/results to publications;
- Application of reasoning and inferencing (e.g., extracting new facts from the integrated data);
- Querying and visualisation of the integrated data.

## 1.3. Cultural Heritage and the Semantic Web

In the last few years, several research projects have focused on cultural heritage content organisation, preservation and integration. For example, the SCULPTEUR project (Addis et al., 2005, Addis et al., 2006, Goodall et al., 2004) provides a dynamic interface to suit the heterogeneous nature of search results related to cultural objects. The MultimediaN E-Culture project (Aroyo et al., 2007, van Ossenbruggen et al., 2007) enables users to explore multiple online cultural heritage repositories via the CHIP browser. These prototype systems aim to improve the discoverability of cultural heritage content via rich metadata.

The Semantic Web (Berners-Lee et al., 2001) promotes interoperability through formal languages and rich semantics. It aims to build a Web where information is exchanged easily between humans and machines. Through a combination of layered standards and protocols for data definition such as the eXtensible Markup Language – XML (Bray et al., 2006), Resource Description Framework – RDF (Beckett and McBride, 2004), the Web Ontology Language – OWL family (McGuinness and Harmelen, 2004), and Uniform Resource Identifiers – URIs (Berners-Lee et al., 2005), the Semantic Web aims to define and expose the semantics associated with data or information in order to facilitate automatic processing, integration, sharing and re-use of the data.

Several research projects have focused on improving the effectiveness of digital libraries in the cultural heritage domain by moving towards a deeper semantic representation of the stored data, through ontologies and semantic annotations. Examples include the CultureSampo (Hyvönen et al., 2009) portal that extended the MuseumFinland ontology (Hyvönen et al., 2006, Hyvönen et al., 2005) and the Archive Mapper for Archaeology – AMA project (Eide et al., 2008, Hernández et al., 2008). The Mellon Foundation also funded six pilot projects: the Master of the Fogg Pietá (Nevin, 2009), the Cranach Digital Archive (Heydenreich, 2009), the Rembrandt Database (Donkersloot, 2009), the Merlin Database (Mellon, 2007), the Raphael Research Resource (Hofmann, 2009) and the Southworth & Hawes Daguerreotypes project (Mellon, 2005). These projects were primarily focused on developing databases for one particular artist or genre, with the aim to integrate all the databases in the final phases of the projects (Oldman, 2010).

Such approaches, although useful, are limited with regard to the discoverability and re-use of the individual components (expressed as compound objects). None of these existing projects have used a common ontology to extract and aggregate knowledge from multiple sources to build a knowledge-base for art/paint conservation and allow inferencing mechanisms across the overall dataset.

Some models have provided the means for describing the resources being dealt with (such as new findings, experimental results and provenance), and enabling knowledge capture to be carried out collaboratively in highly distributed network environments. One example is the Conceptual Reference Model developed by the International Council of Museums' International Committee for Documentation – CIDOC-CRM (Crofts et al., 2010) that provides top-level classes as well as the classes and properties required to capture the provenance information about a painting and its condition state as well as the conservation/preservation activities that it has undergone. A second example is the OreChem project (Lagoze, 2009) that models chemical compounds, chemical reactions and experiments. A third example is the Open Archive Initiative – Object Reuse and Exchange – OAI-ORE project (Lagoze et al., 2008) that models digital objects as aggregations of Web resources.

A number of previous efforts have applied such models to capture semantic knowledge from disciplinary sources. Borkum et al. (2010) and Theodoridou et al. (2010), for example, used OAI-ORE to extract chemicals from chemistry publications. In addition, an extension of the CIDOC-CRM ontology that was able to capture the modelling and query requirements regarding the provenance of digital objects was proposed in Theodoridou et al. (2010).

Similar to these approaches, this research project plans to extract and represent the semantics of unstructured scientific publications in a form that will facilitate re-use and discovery. It is however unique in that it will focuses on the key concepts associated with art/paint conservation (e.g., painting, paint, artist, genre, pigment, chemical, treatment, characterisation, and deterioration mechanism). The proposed services will enable tagging of publications with these core tags. The extracted information will be stored in an RDF triple store where it can be searched and re-used by art conservators.

## 1.4. Aims and Objectives

The aim of this research is to develop and apply the latest information integration, data management and Web 2.0 technologies to collaboratively build a distributed online knowledge-base for 20$^{th}$ century art/paint conservation. Based on the requirements identified from the APTCCARN member meetings (Section 1.1.1), the principal objectives of this research project are:

- To develop an ontology for the preservation of art/paint to develop, curate and share controlled vocabularies to support the evolving knowledge in the art history and conservation science domains. Specifically, the goal of the development of the ontology is to bridge the gap between the physical (e.g., deterioration, condition assessment, exhibition, movement, and treatment) and digital (e.g., paint material, characterisation, physical/chemical structure, and degradation mechanism) provenance of paintings in order to build a comprehensive body of knowledge from existing and emerging preservation techniques;
- To identify the best data models and approaches for aggregating data and sources (e.g., OAI-ORE), capturing both the physical and digital provenance of artworks, and linking multi-disciplinary ontologies (e.g., CIDOC-CRM and OreChem);

- To build services (based on the ontology) that will:
  - Enable conservators and materials scientists to document and describe their own experiments and upload their experimental data to the knowledge-base so it can be shared and re-used efficiently;
  - Enable conservators and materials scientists to automatically extract structured data about past research from relevant publications and websites on art conservation, and to ingest the data into the knowledge-base to enable fast, easy access to and comparison of related cases;
  - Extract related knowledge from key databases (e.g., the Winsor & Newton 19th Century Archive (W&N, 2009), the Infrared and Raman Users Group Spectral Database – (IRUG, 2010), the Dictionary of Australian Artists Online – (DAAO, 2010), the Forbes Pigment Database (MFA-Boston, 2010) and the Paint and Ink Formulations Database (Flick, 2005)), and aggregate it (the knowledge) to the knowledge-base for a seamless federated search over the critical information for art history and materials science;
  - Apply semantic inferencing (e.g., OWL-DL) on the integrated knowledge to precisely extract new facts from the data integration;
  - Enable conservators to search (based on underlying SPARQL service) and visualise data across multi-disciplinary data sources in order to answer art-related queries such as “What solvents will remove surface varnish from the painting *Epiphany*?”

## 1.5. Hypothesis

The primary hypothesis in this thesis is that Semantic Web technologies can provide an effective approach for establishing a collaborative distributed knowledge-base and decision support platform for art conservators. More specifically, the hypothesis is that:

- OAI-ORE compound objects (or RDF graphs) based on an underlying CIDOC-CRM for museum artefacts and OreChem for materials chemistry will provide an effective way to link the different events, activities, objects and agents that are distributed over the Web, and to record the provenance of both the physical and digital artefacts associated with a particular work of art;

- The development of an ontology for art conservation that is based on a common upper ontology will facilitate the integration of relevant knowledge sources to support the search requirements for art conservators;
- The application of semantic tagging tools to art conservation publications will expedite the extraction of machine-processable and re-usable knowledge from full text documents.

More detailed research questions that are tackled include:

- Can a comprehensive knowledge-base comprising RDF graphs be built to support art conservators' information requirements?
- What is the quality of the data model – including the upper ontology, provenance ontology and other ontologies – for underpinning the knowledge-base?
- What sub-disciplinary ontologies exist or need to be developed and incorporated?
- Do existing data models (e.g., CIDOC-CRM) support the requirements of this project or do they need to be extended or refined?
- Is there an existing ontology for describing art deterioration, preservation and conservation concepts?
- If not, are there existing controlled vocabularies that can be re-used to describe artists' materials, paints, painting terminology, conservation terminology, preservation terminology (e.g., techniques, materials and instruments)?
- Can experimental data (samples, experimental processes, observations/measurements, characterisations) be captured and stored in a standardised machine-processable format?
- How accurate is the structured knowledge (that conforms to the ontology, and that is extracted from relevant publications, to enable the re-use, integration and comparison of emerging, current and past knowledge)?
- How efficient and accurate can a large corpus of RDF graphs (derived from publications, related databases and experimental data) be for aggregating, searching, browsing and retrieving (via SPARQL) conservators' information?
- Can semantic inferencing and reasoning (e.g., OWL-DL) be enabled across the RDF graphs in order to extract previously unknown knowledge?

- Can publications be linked to raw and derived experimental data using RDF graphs?
- How can the improvements and benefits of such data models and services for the art conservation community be evaluated?

## 1.6. Approach

To address the above challenges (and requirements identified by the APTCCARN members), this research involves a number of phases that will establish a comprehensive test-bed for evaluating the services that were developed. These include:

- The design and development of an Ontology of Paintings and PReservation of Art – OPPRA as follows:
  - Describing and modelling the information (classes, properties and relationships) of relevance to painting conservators – painting, acquisition, provenance, deterioration, material, physical and chemical processes, treatment, experiment and characterisation;
  - Drawing on existing ontologies that describe art history (e.g., CIDOC-CRM), the physical and chemical properties of materials (e.g., OreChem), and resource aggregations (e.g., OAI-ORE);
  - Drawing on existing controlled vocabularies that include classes and relations not described in the re-used ontologies, such as the deterioration mechanisms and preservation methods from the Getty Art and Architecture Thesaurus – AAT and AICCM Visual Glossary (e.g., darkening, blistering, buckling, cleaning, inpainting, reframing, reweaving and retouching), artistic techniques from the AAT and the International Network for the Conservation of Contemporary Art –INCCA Database for Artists' Archive (e.g., brushwork, sketching and underpainting), and materials and chemicals from the IRUG Spectral Database, CAMEO and the US National Institute of Standards and Technology – NIST Chemistry WebBook (e.g., pigment, paint, oil, bleach, mineral spirit);
  - Extending and refining the employed classes and relationships as required, including the relationships between: paintings and genres, paintings and artists, paintings and samples, paintings and movements, samples and

materials, materials and experiments, characterisation techniques and instruments, instruments and characterisation data, preservation techniques and materials, chemical properties and condition states, etc;

- Evaluating the applicability of the ontology to the offered services (e.g., experimental data capture, structured data extraction from publications, overall knowledge-base, linking experiments to publications, and data aggregation and linking interface) within the context of the 20th Century in Paint project.

- The Design and development of the knowledge-base as follows:
  - Providing the conservation community with a secure Web portal with different levels of collaborative access to data, models, services and storage regarding industrial paint (as illustrated in the next four steps);
  - Integrating data from the provided tools/services into one central repository in a form that complies with the proposed ontology (the OPPRA-based RDF triple store);
  - Enabling semantic inferencing (e.g., OWL 2 RL) over the OPPRA-based RDF triple store to extract new knowledge that is not explicitly mentioned within the aggregated sources;
  - Evaluating its facts by comparing them to a ground truth (e.g., manually assessing the correctness of randomly chosen triples to calculate precision against the actual facts inferred by their sources/sentences).
- The design and development of a collaborative experimental data repository as follows:
  - Implementing a Web-based collaborative workflow system that enables collaborators within the 20th Century in Paint project to quickly and easily describe and publish their experiments and data, the ability to attach permissions and Creative Common Licences to objects, and to search, visualise and compare provenance data (e.g., art history of paintings, experiments, and treatments);
  - Capturing the information in a form that complies with the OPPRA ontology;
  - Enabling the linking of experiments to past case studies (publications and experiments conducted by others);

- Evaluating the effectiveness of linking the experimental data to past publications based on precision and recall.

- The development of text mining tools to extract structured knowledge from past publications as follows:
  - Acquiring a corpus of publications about paint conservation;
  - Developing and employing a Named Entity Recognition – NER service using the OPPRA ontology as the underlying gazetteer, as well as the machine learning approach to resolve ambiguities;
  - Developing and employing a Relationship Extraction – RE service using a rule-based approach to pre-process sentences and extract OPPRA's relations from noun and verb phrases, as well as the machine learning-based approach to extract OPPRA's relations from the pre-processed sentences;
  - Developing a Web-based user interface that enables users to interactively review, correct and refine extracted triples for the accurate capture of structured knowledge;
  - Saving the structured data in the OPPRA-based RDF triple store;
  - Evaluating the performance of the NER and RE tasks by calculating the precision, recall and F-measure.
- The development of a SPARQL search interface to provide access to the distributed, heterogeneous knowledge captured (via the experimental data capture, text analysis, data capture from the external databases and semantic inferencing) as follows:
  - Populating the OPPRA-based knowledge-base with RDF instances from internal databases (Sidney Nolan Paint Archive and Mecklenburg Samples), unstructured information from past publications (*text2triples*) and public databases (e.g., the W&N, DAAO, IRUG and CAMEO);
  - Implementing a user interface that seamlessly converts users' queries to SPARQL queries and returns results with their data sources, URLs to the specific records or sentences, and possible visualisation links depending on the nature of these queries and results.
  - Evaluating the SPARQL search interface based on its performance (i.e., for a given set of multi-disciplinary queries, compare its document and segment

retrieval against the keyword-based search offered by Solr (the open source enterprise search platform from the Apache Lucene project), and on its usability (i.e., deploying it within a team of 20th Century in Paint conservators and scientists to assess its results and functionalities).

## 1.7. Original Contribution

The research presented in this thesis and the research outcomes described within it, make the following original contributions:

- The first ontology (OPPRA) to support the information integration and analysis requirements of art/paint conservators;
- The OPPRA-based knowledge-base to support the storage of experimental data, structured data (extracted from publications) and external databases – required for informed decision-making by the art/paint conservation community;
- A framework and set of services to support the capture, publishing, linking and searching of experimental data associated with the art/paint conservation (based on the OPPRA ontology);
- A set of text analysis tools (a GATE pipeline comprising NER and RE tasks) to support the structured data extraction from publications about art/paint conservation (based on the OPPRA ontology);
- An interface (comprising OWL 2 RL inferencing, SPARQL search, and visualisation) to provide responses to complex multi-disciplinary queries about art/paint conservation, by integrating (and reasoning across) data from relevant existing databases, experimental datasets and publications.

## 1.8. Thesis Outline

The remainder of this thesis is structured as follows:

**Chapter 2** examines the previous, related work in the fields of digital humanities and cultural heritage. The technology, tools and approaches described are designed for the management, analysis and assimilation of historical and digital data on museum artefacts. This chapter also discusses related technologies for knowledge mining and the aggregation of multi-disciplinary data.

**Chapter 3** presents case studies from the 20th Century in Paint project. These case studies were used to define the project requirements, system design and examples for the various models and services proposed throughout this thesis.

**Chapter 4** presents the proposed ontology for art/paint preservation, OPPRA, and discusses its application for meeting the key requirements of the cultural heritage and chemistry informatics domains (e.g., experimental data capture and structured data extraction from past publications).

**Chapter 5** describes the overall framework of the ontology including the requirements, specifications, design and knowledge-base. The framework enables paint conservators to improve their understanding of paint degradation processes, and to identify and document new methods for stabilising, protecting and repairing our valuable but vulnerable paintings. It also discusses the reasons behind the design choices and the technical challenges that the framework must overcome.

**Chapter 6** presents a Web-based platform to enable art conservators and materials scientists to store, search, retrieve, link and visualise the experimental data. The platform captures the users' data (experiments, characterisations, calibrations and data outputs) in a standardised machine-processable format, and links these experiments to past publications and case studies.

**Chapter 7** presents a Web-based platform that performs automatic NER and RE tasks in textual publications about art/paint conservation. The platform extracts standardised machine-processable knowledge or hypotheses from relevant publications so they can be linked, searched and re-used. This interface enables users (particularly conservators and scientists) to add and modify results for the accurate capture of information.

**Chapter 8** describes and evaluates the Data Aggregation and Linking Interface – DALI over the critical information for art history and materials science existing in the 20th Century in Paint project databases, public databases and unstructured information from past publications.

**Chapter 9** is the concluding chapter. It summarises the work done in this research, describes its contributions to the field and draws conclusions from the findings.

# Chapter 2

# Related Work

## 2.1. Introduction

During the recent decades, the contribution of scientists to conservation work related to cultural heritage has grown rapidly. The knowledge in conserving a work of art is not limited to the historical and semiotic analysis. Nowadays, conservation requires a deep knowledge of materials science and nanotechnologies since it is not possible to prevent all natural aging of works of art (Baglioni et al., 2003). Thus, chemists and physicists can contribute greatly to the “controlled death” of works of art because they can provide useful and reliable predictions of the degradation of these works of cultural heritage.

Discovery in the area of art/paint preservation is, however, inefficient. Generally, practitioners use keyword-based searches, navigating and refining results to improve the search accuracy (Elsayed et al., 2011). The well-structured management of documentation is the critical prerequisite for dissemination and sharing, as concluded in the meeting between representatives from over a dozen major museums in the United States and United Kingdom (including museum directors, curators, conservators and scientists) at the Metropolitan Museum of Art in New York on April 27, 2006, who engaged in a frank dialogue regarding the current state of conservation documentation (Rudenstine and Whalen, 2006):

> As the meeting concluded, unanimous agreement was expressed that the digitization of conservation documentation and the sharing of such information among conservators, scientists, museum curators, art historians, and other scholars was highly desirable and of vital importance. It was also

> acknowledged that while public access to such information ultimately would be important, the immediate priority should be the development of mechanisms for the exchange of information among professionals, and that effecting change in institutional practice would be essential if these emerging priorities were to be adequately recognized and served.

Once these crucial semantics about the conservation of cultural heritage materials has been organised in an efficient manner, and attached to its corresponding primary and derived data, they can provide deeper insights into studies than could be grasped from publications or technical reports. This chapter provides an overview of traditional and current approaches to data management and access to art/paint conservation knowledge. More specifically, the related work on art/paint conservation databases, art/paint conservation-related projects, ontologies for paint conservation, and Semantic Web applications to the conservation of cultural heritage materials will be reviewed.

## 2.2. Art/Paint Conservation Databases

Most related work in the field of knowledge capture and reasoning for art/paint conservation has focused on databases that capture information about a specific topic, such as artists (INCCA, 1999) , pigments (MFA-Boston, 2010), paint (W&N, 2009) and publications (CHIN, 2010). Examples include:

- W&N (2009) that provides access to digital images of pigments, paint, varnish and oil recipes from the19$^{th}$ Century Archive of Winsor and Newton;
- The INCCA Database for Artists' Archives (INCCA, 1999) that contains metadata records describing all types and formats of artists' documents (e.g., interviews, technical drawings and installation instructions);
- The Getty Research Institute's Vocabularies (Getty, 2010b) that provide structured vocabularies describing art, architecture, decorative arts, material culture and archival materials;
- The IRUG Spectral Database (IRUG, 2010) that provides a forum for the exchange of infrared (IR) and Raman spectroscopic information, reference spectra and materials.

- The IR-Spectra database (Vahur, 2009) that allows access to a selection of infrared spectra of various paint and coating materials registered at the University of Tartu Testing Centre and Department of Chemistry;
- The Forbes Pigment Database (MFA-Boston, 2010) that provides one central, searchable and readily-accessible location for the Edward Waldo Forbes collection of colorants;
- The Bibliographic Database of the Conservation Information Network (CHIN, 2010) that is the most complete bibliographic resource for the conservation, preservation and restoration of cultural materials;
- CAMEO (MFA-Boston, 1997) that provides a searchable information centre containing chemical, physical, visual and analytical information on historic and contemporary materials used in the production and conservation of artistic, architectural, archaeological and anthropological materials.

All of these databases (and online websites) are designed to provide a specific type of information to art conservators and material scientists. However, they do not:

- Provide services to support the capture, search and retrieval of structured and standardised information describing experiments focused on paint conservation;
- Support the integration of information about artists' techniques, used materials, chemistry of paints, paint degradation processes or paint conservation/preservation methods;
- Support the extraction of structured knowledge about paint conservation from publications;
- Support complex ontology-based queries about paint conservation across distributed databases (e.g., what are the factors that accelerate the occurrence of lead soaps in paintings by Ian Fairweather?);
- Support reasoning across distributed databases using ontology-based reasoning.

## 2.3. Related Projects

Modern Internet portals to cultural heritage collections provide access to aggregations of multimedia content from digital libraries using semantic integration services (Baglioni et al., 2003, Mellon, 2007, Roy et al., 2007). The most common

approach for supporting the mark-up process is the use of metadata (e.g., Dublin Core) (Sugimoto et al., 2002). Whilst this may be as simple as a single keyword tag, it opens the door to interoperability in two ways: either by providing standardised fields with content of a known nature, or by drawing on thesauri, wordlists and other knowledge organisation systems. This section provides an overview of some of the key activities conducted, and databases developed, to support the management of conservation documentation, in particular, the Metropolitan Museum's conservation documentation, the Master of Fogg Pieta online research resource, the Raphael Research Resource, the British Museum's Merlin database, the Lucas Cranach image database, the Rembrandt Database, and the Daguerreotypes of Southworth & Hawes project.

*Metropolitan Museum Conservation Documentation*

A one year survey of the Metropolitan Museum's collection of conservation documentation was conducted to get a clear sense of the scope, methodologies and format of the documentation process, prior to the implementation of six Mellon-funded projects (Green and Mustalish, 2009). The survey collected information on the following aspects within the museum:

- Users of digital documentation (e.g., managers, curators and conservators);
- Physical or digital backups of data (e.g., archiving, storage and locations);
- Types of documents generated by conservation, preservation and scientific activities (e.g., texts of examination records, treatment reports, analytical results and accompanying images in digital format);
- Collection management systems used in the digitising process, with dates, country of origin, history and information regarding media, loans, exhibition history and environmental requirements;
- All information about the museum's cultural objects, including curatorial, conservation, transit, loans, registration and provenance information.

*The Master of Fogg Pieta*

An online research resource to investigate the oeuvre of the 14$^{th}$ century Florentine painter (the Master of Fogg Pieta) (Nevin, 2009) was created by the Courtauld Institute of Art. The project facilitated the Master's study of style, techniques and

materials, as well as the proposal of virtual reconstructions of polyptychs attributed to this artist. The aims of the project were to: foster exchange and scholarly research by concentrating on various paintings by the Master in museums, institutes and private collections in Europe and the United States; and create the Master of the Fogg Pieta/Maestro di Figline Project website to emulate and facilitate the experience of gathering the relevant paintings, and the conservators and curators who study them, in the same place. The materials gathered on the site included new high resolution images of the paintings (IR, visible and X-radiographs), selected analyses of pigments using a range of non-destructive techniques, as well as micrographs from cross-sections and data associated with the analysis of binding media.

*Raphael Research Resource*

The Raphael Research Resource (Hofmann, 2009) is a remotely accessible database created by the National Gallery of London to record a comprehensive range of image and text-based documents (e.g., conservation-derived, technical and art-historical works) by Raphael. The aim was to enrich the resource by incorporating related materials through the collaboration of art institutions to include additional works by Raphael.

*Merlin Database*

The British Museum's wide collection of science and conservation was integrated into the Merlin database (Mellon, 2007). In 2007, information in the Merlin database was made available to share most of the museum's conservation and science documents.

*Lucas Cranach Image Database*

An image database that focuses on the work by Lucas Cranach (c.1472-1553) and his workshop was developed by the Getty Museum (Heydenreich, 2009). The database aimed to provide access to art historical, technical and conservation information on paintings by the artist (more than 700 paintings including 8000 images and documents from 92 contributing institutions), and allow users to make close comparisons of high resolution images on-screen to gain a deeper

understanding of the artist's work and to catalogue his widely dispersed oeuvre of paintings, drawings and prints.

*Rembrandt Database*

An inter-institutional research resource for information and documentation on paintings by Rembrandt was developed in museums around the world in a pilot project funded by the Mellon Foundation (Donkersloot, 2009). The Rembrandt Database is open to anyone, but focuses on academic and museum professionals and students. The aim of the database was to provide a platform for sharing in-depth art historical information, conservation history and technical documentation (e.g., high resolution images with metadata descriptions).

*Daguerreotypes of Southworth & Hawes*

A detailed condition monitoring of daguerreotypes in the exhibition "Young America: The Daguerreotypes of Southworth & Hawes" (Mellon, 2005) was developed in a collaboration project between George Eastman House, the Metropolitan Museum of Art, and the Museum of Fine Arts, Boston to demonstrate alarming changes in the condition of the objects during the exhibition period. The following activities were conducted by the three participant institutions:

- Thirty daguerreotypes (representing a variety of conditions and housing histories) were selected from approximately 1,500 Southworth & Hawes daguerreotypes in the participants' collections;
- In addition to the textual data that were continuously generated for each daguerreotype, large format high-resolution images of the objects captured and conveyed the most useful condition information (e.g., pitting and tarnish). New imaging methods were developed using scanners and microscopy for the accurate documentation of daguerreotypes by reconfiguring an Epson 1640XL scanner. The inverted scanner allowed space for the safe placement of daguerreotypes under the scanner, providing the means to document whole plate daguerreotypes with a non-contact imaging approach;
- Standardised terms to describe the condition of the daguerreotypes based on the terms used in the "Young America" exhibition documentation were used.

Twenty-five damage terms were provided, many with synonyms, descriptions of the terms, and example images of damage types;

- A custom database and a prototype image/information-sharing application for use among the three institutions were developed using the Google Maps Application Programming Interface – API and open source resources such as MySQL, the CASA Image Cutter and several Web-based programming languages including JavaScript and VBScript;
- During the initial survey phase, access to the data via a shared resource was limited to the three participant institutions. Upon completion of the project, the data were to be made available to all interested parties via the World Wide Web.

In addition to the above-mentioned projects, the majority of software applications in this domain are semantic Internet portals that function as delivery channels in various organisations (Baglioni et al., 2003, Reynolds et al., 2004, Hyvönen, 2009), providing a global view of heterogeneous, distributed document materials. Such cultural portals (known as information portals) aggregate either content itself or content metadata only and, thus, they provide effective publication channels and different global search services to end-users.

These systems, however, aim to improve the discoverability of cultural heritage content via rich metadata. They do not draw on semantically-related models that:

- Enable accurate descriptions to entities and relationships among these entities;
- Aggregate data and metadata from diverse and multi-disciplinary domains such as art history, chemistry and material informatics;
- Enable semantic inferencing to extract new facts from the aggregated data;
- Enable complex queries to be performed across multiple domains.

## 2.4. Ontologies for Paint Conservation

The Semantic Web (Berners-Lee et al., 2001) promotes interoperability through formal languages and rich semantics. It aims to build a Web where information is exchanged easily between humans and machines. Through a combination of layered standards and protocols for data definition such as XML (Bray et al., 2006), RDF (Beckett and McBride, 2004), the OWL family (McGuinness and Harmelen, 2004),

and URIs (Berners-Lee et al., 2005), the Semantic Web aims to define and expose the semantics associated with data or information in order to facilitate automatic processing, integration, sharing and re-use of the data.

A number of existing data models/ontologies provide the means for describing the type of resources being dealt with here (e.g., experiments, experimental results and provenance) and enable the capture of knowledge in highly distributed network environments. This section provides an overview of the pre-existing models/ontologies that have been adapted for the paint conservation domain, namely, the CIDOC-CRM, OAI-ORE (including Named Graphs), and OreChem.

### 2.4.1. CIDOC Conceptual Reference Model

The CIDOC-CRM (Crofts et al. 2010) is a formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information. It was developed by interdisciplinary teams of experts, coming from fields such as computer science, archaeology, museum documentation, history of art, natural history, library science, physics and philosophy under the aegis of the International Committee for Documentation of the International Council of Museums.

The CIDOC-CRM was intended to cover the full spectrum of cultural heritage knowledge, from archaeology to art history. Currently incorporating 82 entity types and 262 property types, it is remarkably compact and efficient, given its extremely broad scope. It also has an inherently epistemological structure based around temporal ‘events’ in order to deal with the innate uncertainty of information about the past (Doerr, 2003). The greatest challenge in mapping legacy datasets to the CIDOC-CRM however, was the considerable mental leap required of both museum creators and their technical staff to map their datasets to such an abstract conceptualisation. Although CIDOC had had a number of successes in mapping legacy data to the ontology (Crofts, 2004), and encouraging an adaptive approach which restricted and/or extended the ontology, the process generally required extensive collaboration between curators, IT professionals and CIDOC-CRM experts (Addis et al., 2005). The following list provides examples of the low-level abstract classes covered by the CIDOC-CRM:

- E39.Actor/E21.Person/E74.Group – These classes comprise people (either individually or in groups) who have the potential to perform intentional actions. The CIDOC-CRM doesn't model the types of entities that perform these actions in the conservation science domain (e.g., organisation, museum, manufacturer, artist, conservator);
- E77.PersistentItem – This class comprises items that have a persistent identity, that is, they can be repeatedly recognised within the duration of their existence by identity criteria rather than by continuity or observation. The CIDOC-CRM doesn't model the types of persistent entities that exist in the conservation science domain (e.g., artefact, painting, sample, E73.InformationObject extension – below, E57.Material extension – below);
- E73.InformationObject/E31.Document) – These classes comprise identifiable immaterial items (e.g., images, text and multimedia objects). The CIDOC-CRM doesn't model the types of information objects denoting the conservation science domain (e.g., characterisation data, spectra, backscattered electron image, interferogram, X-ray image);
- E57.Material – This class comprises the concepts of materials, denoting properties of matter before its use, during its use, and as incorporated in an object (e.g., ultramarine powder, tempera paste). An extension to this class is necessary to model the relationships between materials in the science/chemistry domain (e.g., OreChem);
- E26.PhysicalFeature – This class comprises identifiable features that are physically attached in an integral way to particular physical objects. An extension to this class is required to model the types of physical features denoting cultural heritage artefacts (and conservation science materials) (e.g., physical attribute, chemical attribute, visual attribute, temporal and spatial attribute, dimension, condition state, physiochemical attribute, colour, age, brightness, brittleness, fragility and dryness);
- E5.Event/E7.Activity – These classes comprise changes of states in cultural, social or physical systems (regardless of scale) brought about by a series or group of coherent physical, cultural, technological or legal phenomena (e.g., changes of state will affect instances of E77.PersistentItem or its subclasses). An extension to these classes is required to model the types of events/activities

that occur in the conservation science domain (e.g., chemical process, reaction, condition change, deterioration, environmental process, conservation, cleaning, acquisition, experiment, characterisation, production and painting process);

- E1.Entity – This class comprises all things in the universe of discourse of the CIDOC-CRM. An extension to this class is required to conceptualise other types of entities that cannot be (directly) specified using the CIDOC-CRM for the conservation science domain (e.g., genre, technique, artistic technique, characterisation technique and inscription).

### 2.4.2. Open Archives Initiative Object Re-use and Exchange (OAI-ORE) vs. Named Graphs

OAI-ORE (Lagoze et al., 2008) is an international collaborative initiative focusing on a framework for the exchange of information about digital objects between cooperating repositories, registries and services. OAI-ORE aims to support the creation, management and dissemination of new forms of composite digital resources and to make the information within these objects discoverable, machine-readable, interoperable and re-usable. OAI-ORE defines the following key entities:

- **Aggregation**: is a resource type, which is a set of other resources (e.g., Publication);
- **ResourceMap:** a Resource Map has a URI and a machine-readable representation that provides details about the Aggregation. It lists the resources that are part of the Aggregation (via *ore:aggregates*) and expresses the relationships and properties pertaining to all these resources, as well as metadata pertaining to the Resource Map itself (e.g. who published it and when it was most recently modified);
- **AggregatedResource**: is an Aggregation (resource) that is constituent of another Aggregation. Both aggregations (Aggregation, and AggregatedResource) are connected via the relation "*aggregates*" (e.g., Book *aggregates* Chapter).

Named Graphs (Carroll et al., 2005) are endorsed by the OAI-ORE initiative as a method of publishing compound digital objects that clearly states their logical boundaries (Lagoze et al., 2008). The Named Graphs method does this in a way that

is discipline-independent, but that also provides hooks to include rich semantics, metadata, ontologies and rules.

Named Graphs do not, however, propose any particular conceptual model or ontology. What is called a “Named Graph” is very simplistic and cannot be considered as a conceptual model (it resembles more a container for core knowledge triples). In brief, it states that each triple (subject, predicate, object) can be associated with representation information (i.e., information needed for interpreting, grouping or describing the core knowledge). This may include information about the structure and the semantics of the core triple.

An overlap between OAI-ORE and Named Graphs is necessary to describe the knowledge obtained for the art preservation domain (i.e., art history, and materials science). This semantic overlap between digital resources is needed for the following reasons:

- Using OAI-ORE to represent the core triples for the art preservation domain would result in triples’ components (i.e., subjects, predicates and objects) being transcribed with separate relations to each statement (e.g., Statement1 hasSubject Subject1; hasPredicate Predicate1, hasObject Object1). Drawbacks of this method include:
  - The size magnification of the knowledge-base – resulting in poor performance;
  - Complex inferencing applications – due to the complex assignments of triples;
- Using Named Graphs to contain the core triples would result in poor transitive inferencing associated with the digital resource (e.g., Publication aggregates Sentence, Sentence aggregates Triple → Publication aggregates Triple). This representation can be easily achieved using OAI-ORE, but would be impossible using Named Graphs (without adding metadata to the digital resource to indicate whether a transitive relationship exists between publication, sentences, triples, as well as databases, records and triples).

### 2.4.3. OreChem

OreChem is a model developed by the OreChem project (Lagoze, 2009) for research and dissemination of scholarly materials in the chemistry community based on OAI-ORE. The OreChem project is a Microsoft-funded collaboration between Cornell, University of Cambridge, University of Indiana, Penn State, PubChem and the University of Southampton. The collaborators are designing a graph-based object model for the chemistry domain that is built around the central role of the “molecule” and the “chemical compound” and the underlying specifications of OAI-ORE. However, the OreChem model is only focused on the following inorganic crystallography models:

- Chemistry domain ontology – This is currently a small ontology that clarifies some fundamental relationships in the chemistry domain. Key concepts in this ontology are: chemical element, chemical species, molecular entity and chemical role;
- Properties ontology – This is an ontology of over 150 chemical and materials properties, together with a first set of definitions and symbols (where available and appropriate) and some axioms for the typing of properties;
- Measurement techniques – This is an ontology of over 200 measurement techniques and also contains a list of instrument parts and axioms for the typing of measurement techniques. At the time of writing, it did not include information about minimum information requirements for measurement techniques (e.g., the measurement of a boiling point also requires a measurement of pressure) and other metadata, but this was expected to be added at a later stage;
- Polymers ontology – This ontology contains terms which are in common use across polymer science as well as a taxonomy of polymers based on the composition of their backbone.

## 2.5. Semantic Web Applications for the Conservation of Cultural Heritage Materials

In the cultural heritage domain, a number of previous projects have aimed to develop open-source software applications that address the core needs of the art/paint conservation community using Semantic Web technologies. Chapters 6-8 contain

the specific details of previous related efforts to the problems of art/paint conservation documentation management. These include:

- **Experimental data capture** – related work on this topic is described in Section 6.2;
- **Structured knowledge extraction from past publications** – related work on this topic is described in Section 7.2;
- **Ontology-based data integration capture** – related work on this topic is described in Section 8.2.1;
- **Ontology-based reasoning and querying for art conservation** – related work on this topic is described in Section 8.2.2.

## 2.6. Summary

This chapter provided an overview of existing approaches to the capture, storage, integration, sharing and re-use of data within the art/paint conservation domain. A summary of these approaches is provided below.

- Approaches that use traditional databases only provide a specific type of information to art conservators and materials scientists (e.g., artists, artefacts, materials, characterisation images);
- Some approaches focus on cultural heritage content organisation, preservation, and integration via rich metadata. Such approaches do not draw on semantically-related models that: a) enable accurate descriptions of entities and relationships; b) aggregate data and metadata from diverse and multi-disciplinary domains, such as art history, chemistry and material informatics, via a common ontology; c) capture domain expert knowledge within an ontology; d) enable semantic inferencing to extract new facts from the aggregated data; and e) enable complex queries to be performed across multiple domains;
- Some approaches focus on improving the effectiveness of digital libraries in cultural heritage by moving towards a deeper semantic representation of the stored data through ontologies and semantic annotation. Such approaches (described in Chapters 6-8) do not focus on key concepts associated with art/paint conservation which would enable structured knowledge capture and the linking of experiments and publications based on common ontology for paintings and art preservation.

# Chapter 3

# Case Studies

## 3.1. Introduction

This chapter provides three case studies that illustrate the requirements of art conservators who need answers to questions relating to art conservation, curatorial practice and paint analysis/composition/chemistry. These case studies provide information about the tools and strategies that teams of the Australian Research Council-funded project called "the 20th Century in Paint" used to address 20th century art conservation issues in Asia and the Pacific. The following three case studies were selected:

- The 20th Century in Paint as the overall framework that underpins the various research activities conducted by the project collaborators and represents the typical workflow of art conservation researchers;
- Sidney Nolan's experimentation with commercial materials as an example of the study of historical development of the Australian housepaint for the identification of the range of likely products used by Australian artists in the 20th century;
- Zinc oxide-centred deterioration of modern oil painting materials as an example of the identification of zinc soap aggregates in association with the deterioration of paintings.

## 3.2. The 20th Century in Paint

Throughout the 20th century, artists enthusiastically embraced new materials in ways that radically changed the art process. Choices offered by traditional pigments and media were extended by technological developments that saw the introduction of synthetic media, new pigments and dyes, and additives that modified paint handling

and performance. This resulted in revolutionary art practices in Australia and across the Southeast Asia-Pacific region, but occurred with a lack of understanding of the preservation issues associated with such usage. Today collectors, curators and conservators are confronted with material-based preservation questions (Section 3.5), but lack the sustained and integrated research base to inform their decision-making.

The 20th Century in Paint project utilises expertise and collections from ten public institutions and four universities, namely, the University of Melbourne, University of Queensland, Art Gallery of New South Wales, National Gallery of Victoria, Artlab Australia, Queensland Art Gallery, Getty Conservation Institute, Tate Britain, National Gallery of Malaysia, J. B. Vargas Museum at the University of the Philippines, Silpakorn University in Thailand and the Southeast Asian Ministers of Education Organisation's Project in Archaeology and Fine Arts (20thcpaint, 2010b).

The project aims to provide fundamental information that answers questions relating to conservation, curatorial practice and paint manufacture. It examines new media, pigments, dyes and additives that led to the creation of revolutionary works of art in the 20th century in both Australia and Southeast Asia. It considers how these materials impact on the paint handling, performance and permanence of modern art. In turn, the project aims at improving the understanding of the scientific analysis of cultural materials, filling critical gaps in understanding the effect of diverse climates on artworks, informing the preservation of modern art, and developing interdisciplinary data management systems.

The project program is interdisciplinary and four teams tackle distinct but integrated research programs. Together, they provide unique industry-related research in conservation, art history, e-humanities, curatorial studies, physical and biochemical sciences. These research programs are:

- Art history and conservation (team 1);
- Material developments and deterioration (team 2);
- Scientific tools and techniques (team 3);

- IT tools and techniques (this work as part of team 4 illustrated throughout this thesis).

Activities conducted within the 20th Century in Paint project helped to understand the nature of the problem (and the questions relating to conservation, curatorial practice and paint manufacture), and to define conservators' and scientists' requirements (which further defined the requirements and aims of this thesis). To illustrate how research activities are conducted within the 20th Century in Paint project, a typical workflow of art conservation research is provided in the following list:

1. The art conservator starts with the process of identifying a painting and any issues/problems that the painting has (e.g., cracking, discolouration, peeling, mould, fading, swelling, and tearing);
2. If a problem is identified, then a documentation process starts (before any treatment) by writing a condition report with information on the problem, observations, images (if available) and reasons (if known);
3. Further analysis/characterisations are usually needed to support these observations, and to identify the precise cause of the problem (e.g., materials, chemical processes, physical pressure, water, humidity and radiation). Each characterisation process (e.g., FTIR, SEM, TEM, X-ray, XRD, XRF, PY-GC-MS and Synchrotron Radiation) usually consists of sub-tasks (independent workflow) that may be documented separately, but linked via the documentation process (step 2). For example, FTIR characterisation may consist of taking (or re-using) a sample from an invisible area on the painting, placing the sample under a light beam, measuring how much of the light is absorbed using different wavelengths, documenting the calibrations employed in the characterisation process, saving any data derived at any stage of the process (e.g., interferogram, images, and spectrographic data), and writing observations during the characterisation process;
4. Art conservators then try to analyse these results and read publications (journal articles, websites, Wikis, databases) about related problems to identify one or more possible approaches to treating the problem(s) of concern;
5. Experiments (e.g., controlled temperature, relative humidity, paint mixtures, support materials, brushwork methods, and colour/radiation variations) on mock

samples (or on invisible areas on the painting) are then conducted for each possible treatment;

6. Samples (taken from the treated areas, or modified samples from previous experimentations) are re-analysed with the same characterisation techniques (step 3) to check if the problem is being removed;
7. Final results are then assessed and documented (separately for each approach, but linked via the documentation process(step 2));
8. If possible, a paper is published about the problem(s), methods, results and final observations.

Figure 3.1 illustrates a visual graph of this workflow.

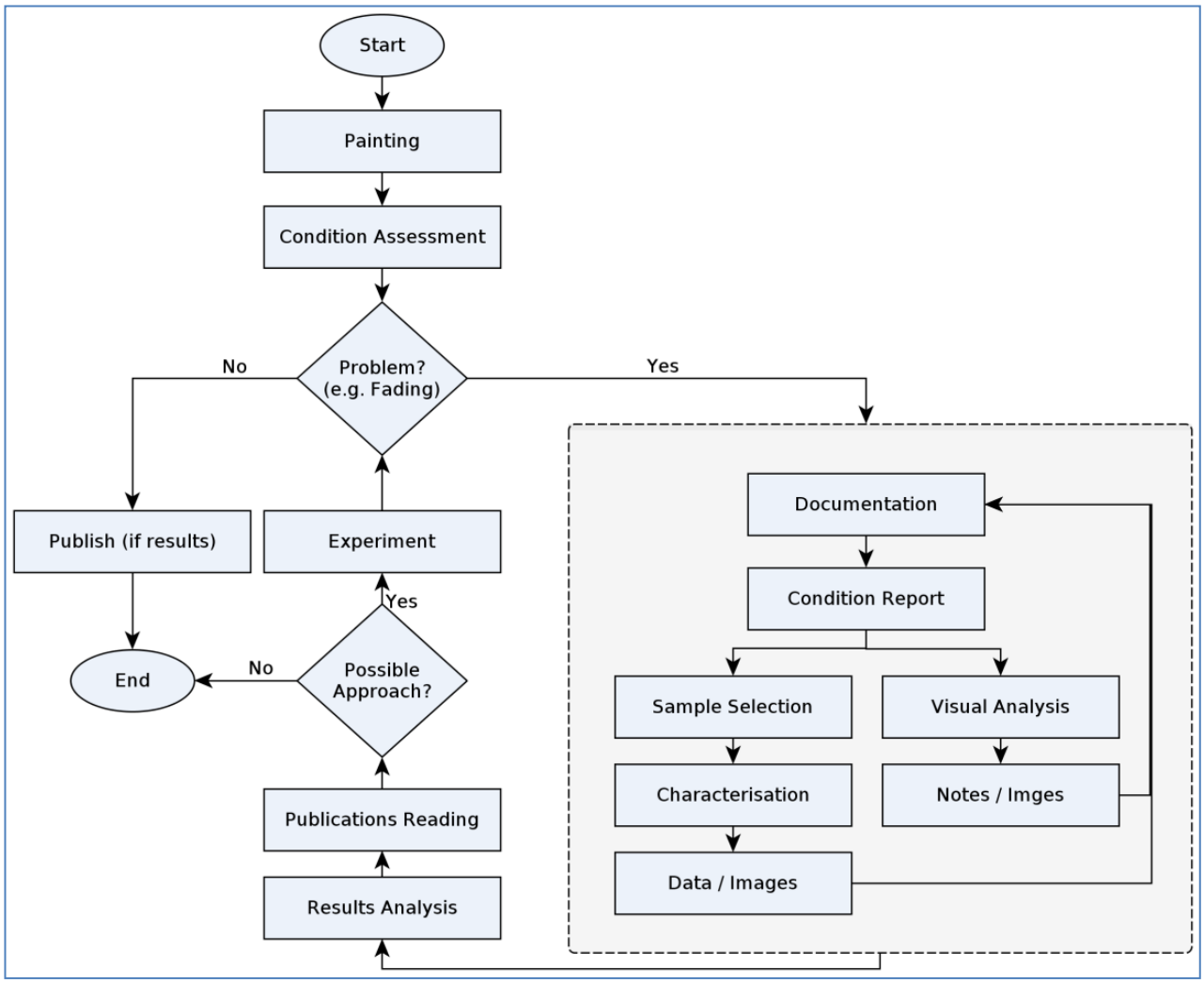



**Figure 3.1: Typical workflow of art conservation research**

The following sections provide details on two case studies within the project's research programs:

- Sidney Nolan's experimentation with commercial materials – conducted by Paula Dredge within team 1;
- Zinc oxide-centred deterioration of modern oil painting materials – conducted by Gillian Osmond within team 2.

## 3.3. Sidney Nolan's Experimentation with Commercial Materials

The research into Australian artist, Sidney Nolan's experimentation with commercial material was conducted by team 1 (Dredge, 2010, Dredge, 2012) in the 20th Century in Paint project. This research aimed to investigate the artist's use of Australian housepaint from the 1920s to 1950s and to extend art conservators' understanding of the complexity of these artistic materials and their components. The following provides a summary of team 1's research, details of which are available from (Dredge, 2010, Dredge, 2012).

- **Background**: Prior research has begun to document the complex history of commercial paint production in the US and UK (Standeven, 2011) and its use by artists; however, detailed studies of the corresponding situation in Australia have not been undertaken. This is partially due to the difficulty in distinguishing housepaints from art paints by their binder and pigment components.
- **Nolan's use of housepaint**: Nolan appeared to have had little regard for traditional artists' paints because he was already practised in the use of locally manufactured housepaint; artists' paints were always expensive relative to housepaints; and during the Second World War, they were also hard to obtain as manufacturers ceased production.
- **Historical study of Nolan's experimentation with commercial paints**: A large amount of historical evidence provided reference material for this research into Nolan's use of house paints. Historical evidence includes:
    - Nolan's' writings and interviews that included descriptions about his works, the mediums used and their physical appearance.
    - A collection of paints, solvents, cobalt naphthenate (manufactured in Australia by Reichhold/A.C. Hatrick Company), and other assorted materials

from Sidney Nolan's studio in Wahroonga, Sydney, that were in use in the 1949-53 period (Figure 3.2).

- Nolan's letters to Sunday Reed (written while in the Australian Army) (Unpublished manuscript (1924-1981), The University of Melbourne).
- A receipt for linseed oil and turpentine purchased from Reichhold/A.C. Hatrick Company (plus handwritten notes, and a notation with street directions to the Botany plant).

- **Findings from the historical study of Nolan's experimentation with commercial paints:**
  - Initial analytical results from the tins of Ripolin from Nolan's Wahroonga studio confirmed that these paints were natural oil-based enamels. This suggested that prior to Nolan's departure from Australia in 1953, the media of his paintings were unlikely to be synthetic polymer paint, as they have often been catalogued.
  - The collection also demonstrated one of the features of Ripolin paint that must have been attractive to Nolan, namely, its availability in a huge range of colours (72 colours, including many that were bright and intense).
  - Synthetic polymer paints were used by Nolan. This is evidenced by his purchase and use of the DUCO and DULUX brand paints (manufactured in Australia by the British Australian Lead Manufacturers) and the Dynamel brand paint (manufactured in Australia by Taubmans).
  - The presence of a lead white-based oil paint (which included a drier) in his Wahroonga studio, suggested that Nolan may have been experimenting with making his own enamel paint.
- **Analytical study of Nolan's materials used in his artworks**: Further analysis of the artist's materials found in his Wahroonga studio showed the following observations:
  - Microscopic analysis of the house paints identified the presence of alkyds and nitrocellulose resins.
  - Analysis of 20$^{th}$century paints, and in particular commercial paints, indicate that the proportion of organic materials to pigments is considerably reversed. Powerful tinting pigments were developed that required less volume of

pigment and glossy oil-based house paints did not need large amounts of bulking agents as they were fairly liquid.

- Infrared beam-line analysis (at the Australian Synchrotron) on samples of Ripolin paint revealed the formation of metallic soaps due to a chemical reaction between the oil and metallic components of the paint. Metal soaps have the potential to be mobile within the dried film and may develop into large aggregates/lumps in the paint film and cause problems due to their solubility when exposed to solvents during cleaning.


**Figure 3.2: Collection of paints, solvents and other assorted materials from Sidney Nolan's studio in Wahroonga, Sydney (in use 1949-53)**

## 3.4. Zinc Oxide-Centred Deterioration of Modern Oil Paintings

The research into zinc oxide-centred deterioration of modern oil paintings has been conducted by team 2 (Osmond, 2012) in the 20th Century in Paint project. This research investigated paintings known to contain, or with the potential to develop, zinc carboxylate aggregates. It aimed to: improve our understanding of the conditions that lead to structural or optical deterioration; and enable the design of appropriate conservation strategies for storage or treatment of vulnerable paintings. The following provides a summary of team 2's research, details of which are available from (Osmond, 2012, Osmond et al., 2012, Osmond et al., 2005).

- **Background**: Prior research findings from across 20th century artworks have suggested that the stability of zinc oxide-containing paints is strongly dependent on the differing fatty acid composition of oils used in paint formulations, and the

variable reactivity of the different types and grades of zinc oxide pigment (which is linked to particle size, shape, surface area and stoichiometry). In addition to these intrinsic properties, additional compositional and environmental factors were found to influence film stability.

- **Housepaint performance in Australian climates**: Mid-20th century zinc-based housepaints were found to have performed poorly in Australian climates in comparison to comparable paints applied in European environments (e.g., as depicted in Figure 3.3, zinc soap aggregates have been identified as a cause of surface deterioration of paintings).
- **The influence on zinc carboxylate distribution**: Several factors were considered in terms of their influence on zinc carboxylate distribution including:
    - The presence of other metal ions from pigments with which zinc oxide is frequently combined; particularly, lead and titanium whites.
    - The relationship between the metal soaps present as starting components and the zinc carboxylate phases forming in situ, with a particular focus on aluminium stearate.
    - The effects of temperature and relative humidity on reaction dynamics in the dried film, including the significance of temperature cycling which is known to favour zinc carboxylate aggregation within liquid paint systems.
- **Primary Focus:** The preliminary work by team 2 focused on the following compositional factors:
    - Historical research into technological developments in zinc oxide pigments and paint production that affect pigment-oil interactions.
    - Analysis of cast films of ‘Control’ and commercially prepared art paints naturally aged for up to 30+ years which incorporated zinc oxide in different oils and were combined with other pigments and additives.
    - Investigation of paint samples from actual paintings affected by zinc carboxylate aggregation.
- **Analytical study of zinc oxide-centred deterioration of artworks**: The characterisation techniques used included the optical and electron microscopy of surfaces and samples in cross-section, SEM-EDX analysis, TEM of thin sections and pigment samples, synchrotron source infrared microspectroscopy of thin sections, and the FTIR and XRD of the bulk film and aggregates. Observations

(used to inform the design of fundamental experiments e.g., specific interactions under controlled conditions) indicated that the sensitivity of affected paint films to pH and solvents commonly used in conservation treatments was an additional question of interest and should be considered in future research by team 2.

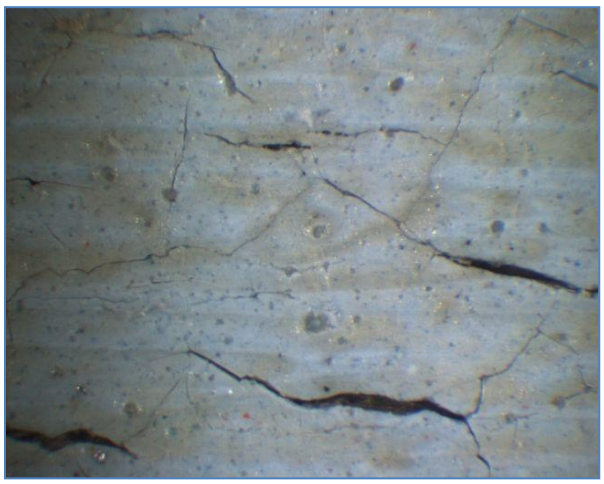


**Figure 3.3: Magnified surface detail of a painting where zinc carboxylate aggregates have erupted through the paint film (Queensland Art Gallery Collection 2010)**

## 3.5. Categorising the Questions Asked by Art Conservators

By analysing these case studies and surveying members of the Asia Pacific Twentieth Century Conservation Art Research Network (APTCCARN), a list of examples of the types of queries to which art conservators need answers were identified. Although this thesis cannot address all the possible specific queries that were identified, it is possible to categorise the queries into three different types, that can be investigated and hopefully supported. These three categories of queries are considered significant in the context of this study because: they are unique to painting conservation; they cover the majority of user requirements; they involve increasing levels of complexity and increasing numbers of linked datasets. The three categories of queries that were identified and examples of queries that belong to each category are listed below:

- Questions regarding the condition changes of materials and/or references made to condition changes in online publications. Examples:
  - List references that report the degradation of lead carbonate;
  - Find case studies that deal with the darkening of chromium yellow;
  - What chemical reaction is involved in the oxidation of cadmium?

- Questions regarding the investigations of materials and their degradation mechanisms (e.g., experimental results or characterisation results). Examples:
  - List all references that describe light bleaching experiments on flake white;
  - List all references that describe the characterisation of aluminium sulphate;
  - List all references that describe FTIR-ATR characterisation of W&N paints;
  - List all investigations that use Portable XRF to characterise zinc oxide after exposure to different levels of humidity;
  - List all references that describe the characterisation of Ripolin paint that contains titanium dioxide;
- Complex, cross-disciplinary questions that involve both art history/provenance (e.g., artists, paintings and materials) and conservation science (degradation mechanisms and treatment activities). Examples:
  - What is the best way to treat zinc oxides occurring in paintings by Rover Thomas?
  - What are the factors that cause or accelerate the occurrence of lead soaps in paintings by Ian Fairweather?
  - What is the best solvent for removing varnish from acrylic paintings that exhibit cracking?
  - Give me all experiments that involve the removal of lead soaps from blanched acrylic paintings with mineral spirits
  - What solvents will remove surface varnish from the painting *Epiphany*?
  - What are manufacturer-included additives contribute to longer-term stability problems?
  - List all paintings that show severe darkening due to heat exposure;
  - List all oil paintings that show cracking due to metal soap formation;
  - List all condition reports that show the presence of lead soap aggregates in artworks painted using Ripolin.

## 3.6. Requirements

In order to support the typical workflow of art conservation research (Section 3.2), and to provide answers to the typical questions asked by art conservators (Section 3.5), it is necessary to develop a Web portal that provides an integration/linking of heterogeneous data sources (represented in a common data model), and a

federated search (or an ontology-based search) interface to these distributed data sources. The specific requirements include:

- Repositories for experimental and characterisation data;
- Structured data extraction from past publications;
- A method to link publications to persistent data underpinning the results in the publication, namely, data that provides evidence to support the theory or hypothesis in the publication;
- Integration of, reasoning over and searching across the distributed data from internal repositories, publications and external databases, in order to provide decision support tools.

### 3.6.1. Capturing Experimental and Characterisation Data

To support the capture and management of the experimental data generated by art conservators and materials scientists in the 20th Century in Paint project, an event-aware framework based on a common, machine-processable model is needed. The required framework should be able to provide the following services:

- Online repositories that enable art conservators and scientists to describe each investigation, paint samples (e.g., name, brand, year, binding medium and location), experimental conditions (e.g., artificial aging, controlled humidity, instruments and calibrations), characterisation techniques (e.g., SEM, FTIR and Portable XRF) and characterisation results (e.g., X-ray images and spectra graphs). The data storage in this phase should also provide:
  - Persistent links (URLs) to continue providing access to art history and materials science resources from publications;
  - Access control mechanisms to protect the results until they are ready to be published;
- A Web-based collaborative workflow system that enables collaborators to quickly and easily describe, edit and publish their experiments and data, with built-in provenance. The workflow system in this phase should also provide:
  - The ability to easily generate different content for different tasks;
  - The ability to add provenance data at any stage of the documentation process;
- The ability to link experiments to past experiences and publications.

### 3.6.2. Extracting Structured Data from Past Publications

To support the linking and comparison of related research outcomes described in publications, a structured data extraction from past publications (represented in a standardised machine-processable format so that it can easily be discovered by the team of conservators) is needed. The required text analysis and structured data extraction framework should be able to provide the following services:

- The extraction of meaningful statements from free-text publications that express facts or accepted knowledge associated with paint conservation;
- A Web-based user interface that enables art conservators to quickly and easily review, visualise and edit results graphically to ensure accurate knowledge capture;
- A knowledge-base of facts about the conservation of paintings that can be easily integrated with additional knowledge captured through further publications, databases and experiments.

### 3.6.3. Integrating, Reasoning and Searching Multi-Disciplinary Datasets

To support the integration and linking of the published data, and to build the knowledge-base of facts about the conservation of paintings, a data aggregation and linking interface based on a common, machine-processable model is needed. The required framework in this phase should be able to provide the following services:

- A data integration system that provides access to historical and provenance data associated with 20$^{th}$ century art/paint conservation. The integrated data needs to be captured from the following data sources:
  - Internal repositories for members of the 20$^{th}$ Century in Paint project (e.g., Sidney Nolan Paint Archive, and Mecklenburg Samples);
  - Textual publications about paint conservation (e.g., JAIC, JSTOR Studies in Conservation, Analytical Chemistry, and AICCM Bulletin);
  - External databases identified as valuable sources by the 20$^{th}$ Century in Paint project members (e.g., W&N Archive, DAAO, IRUG Spectral Database, and CAMEO);
- A linking and reasoning system that infers new facts not explicitly mentioned in the literature. Examples include the relation between condition changes and condition states (e.g., ‘Burned’ *describes* ‘Burning’), the relation between

characterisation techniques and instruments (e.g., 'FTIR Spectrometry' *usesInstrument* 'FTIR Spectrometer'), as well as the relation between creation activities, actors and techniques (e.g., 'The Camp' *hasArtisticTechnique* 'Thick Stroke' and 'The Camp' *isPaintedBy* 'Sidney Nolan' → 'Sidney Nolan' *usesArtisticTechnique* 'Thick Stroke');

- An integrated and event-aware informatics framework/knowledge-base that will:
  - Establish a semantic Linked Open Data network, linking art history, paintings, people, artistic techniques, provenance, samples, experiments, characterisations and preservation treatments;
  - Provide answers to more sophisticated queries than traditional databases that provide silos of information about specific aspects of art conservation (e.g., "Give a list of oil paintings showing cracks due to metal soap formation");
  - Provide a decision support tool to recommend the most appropriate method given a specific art conservation problem (e.g., "What is the best solution for cleaning mould from a 20$^{th}$ century acrylic painting by Ian Fairweather?").

## 3.7. Summary

In this chapter, three case studies for the activities conducted by art historians, conservators and materials scientists within the 20$^{th}$ Century in Paint project were discussed. The case studies provided information about the tools and techniques used to investigate 20$^{th}$ century paintings in Australia and Southeast Asia-Pacific with the aim to inform conservators and art historians about the making of artworks and their ageing characteristics and to assist in the informed decision-making about the artworks' conservation and care.

The case studies enabled the identification of the typical workflow of art conservation research, and the questions that art conservators want answered. The case studies also enabled the identification of the sources of information that, together, would provide answers to these questions.

To assist the data management tasks in the selected 20th Century in Paint project case studies (and as an overall prototype for the conservation science domain), the following requirements were derived:

- Capture new information in a form that complies with a common data model;
- Extract structured knowledge (based on the model) from raw data and text gathered from free-text publications and discussions;
- Allow heterogeneous information to be aggregated, linked, searched and analysed.

The specific details, system design and implementation of the data management framework to fulfil these requirements are presented in Chapters 4 to 8.

# Chapter 4

# Ontology of Paintings and PReservation of ART – OPPRA

## 4.1. Introduction

Painting conservation has evolved into a highly multi-disciplinary research topic that requires the integration of knowledge about art history (artworks, artists, artistic techniques), the physical and chemical properties of paint and pigments, paint conservation and cleaning methods, experimental data and the results of sophisticated characterisation techniques (e.g., SEM, XRD, and Raman spectroscopy) that are used to determine the precise causes of degradation or discoloration.

The challenge is that the relevant data and metadata are highly heterogeneous and distributed across databases, scholarly publications and the Web. Expertise, also, is distributed across art galleries, conservation centres and universities around the globe. Although it is possible to find some concentrated authoritative collections of information on this topic on the Web (e.g., JAIC (COOL, 2002), Smithsonian Museum Conservation Institute (MCI), Getty Conservation and Research Institutes (Getty, 2010a), CAMEO (MFA-Boston, 1997), and Forbes Pigment database (MFA-Boston, 2010)), the information is often embedded within databases or within highly unstructured textual documents and the relevant information is difficult to extract, re-use, interpret, correlate or compare. Moreover, it is often the case that the raw

images or the raw spectrographic data associated with the analysis of a particular painting or paint samples are not accessible via the related publication. For example, the experimental data underpinning publications that describe the long-term effects of different environmental conditions (humidity, temperature, UV light) on different paints are not accessible, verifiable or re-usable.

The distributed, unstructured, heterogeneous nature of the relevant data makes it extremely difficult for conservators to search and aggregate information to find answers to the problems that they face. For example, consider the following hypothetical example. An art conservator at the Queensland Art Gallery recently wanted to know: "what is the best solvent for removing the surface coating from the painting *Epiphany*". The QAG database reveals that *Epiphany* was painted by Ian Fairweather in 1962, and purchased by the QAG in 1984. The DAAO (2010) tells that during this period, Ian Fairweather frequently used Dulux acrylic paints coated with shellac. The CAMEO tells that the best solvent for removing shellac is methyl ethyl ketone. But the process of discovering these different pieces of information and linking them to answer the original question is both extremely time consuming and cumbersome and involves reading through long textual resources (e.g., a biography of Ian Fairweather). ***The hypothesis in this research is that an ontology/or ontologies can be usefully applied to the paint conservation domain to help conservators integrate disparate multi-disciplinary datasets to answer their complex questions.***

Ontologies have been successfully applied in many fields (biomedical (Bundschus et al., 2008), environmental sciences (Raskin and Pan, 2005), literature (Barbosa-Silva et al., 2010), etc.) to enable data integration, knowledge acquisition, semantic annotation and reasoning for knowledge discovery purposes. A survey of the literature was undertaken (Chapter 2) and indicated that there is no existing ontology designed to support knowledge representation and reasoning for painting conservators.

However, there are a number of existing ontologies or vocabularies that provide data models for describing particular aspects of art conservation. For example:

- CIDOC-CRM (Crofts et al., 2010) provides a data model for describing the provenance of artworks;
- The ChEBI ontology (Degtyarenko et al., 2006) describes chemical entities of biological interest;
- ChemAxiom (Adams et al., 2009) provides an ontological framework for Chemistry in Science;
- The Materials ontology (Ashino, 2010) provides a common data model for exchanging information about materials (structure, composition and properties);
- AAT (Getty, 2010b) provides a structured, controlled vocabulary for describing artworks, but does not include support for describing the materials they are composed of (e.g., paint types). For example, in AAT, “oil painting” is a technique, not a material.

The aim is to build an ontology, the Ontology of Paintings and PReservation of Art – OPPRA, that describes the classes/entities, properties and relationships of relevance to painting conservators, by drawing on existing ontologies and vocabularies where available – but also extending and refining them as required.

OPPRA is designed to provide a common, machine-readable formal representation of the knowledge in the domain of art/paint conservation. The aim of OPPRA is to:

- Document and describe experiments conducted by conservators and scientists and allow them to upload their data and findings to the knowledge-base, share and re-use this data among them, and make it accessible publicly;
- Extract structured data about past research and experiments from relevant publications;
- Provide a data integration and linking interface that aggregates information and reasons over the extracted knowledge from internal and external datasets – that were identified as invaluable sources for art conservation and material science domains (e.g., W&N, DAAO, IRUG Spectral Database, CAMEO, Forbes Pigment Database, Color of Art Pigment Database, FT-IR Spectra of Binders and Colorants, NIST Chemistry WebBook, and Paint and Ink Formulations Database).

## 4.2. Development of OPPRA

The aim of the OPPRA ontology is to develop, curate and share controlled vocabularies to support the evolving knowledge in the art/paint conservation domain. Specifically, OPPRA's goal is to document descriptions of physical artefacts (e.g., genres, condition states, artists, artistic techniques), events (e.g., condition assessments, exhibitions, movements, treatments), deterioration mechanisms (e.g., discolorations, oxidisations, damages), and related digital information (images, characterisation data, spectrographs, publications) associated with art/paint conservation – in a standardised, re-usable, and machine-processable format.

The aim is to design the OPPRA ontology by drawing on previous related work undertaken in the cultural heritage and chemistry informatics domains, where possible. OPPRA is decided to be based on an upper and advanced knowledge representation system – leveraging existing peer-reviewed ontologies and vocabularies for a number of reasons. These reasons were:

- To deal with the technical aspects of ontology construction easily and reliably;
- To avoid ambiguous interpretation and enable compatibility between the concepts from OPPRA and other domains;
- To provide a library of richly structured and well-understood abstract data types;
- To enable integration of high priority datasets to serve a community of conservation practices.

The OPPRA ontology extends the following sub-ontologies:

- CIDOC-CRM (Crofts et al., 2010) that provides the top-level classes as well as the classes and properties required to capture the provenance information about a painting, condition state, as well as the conservation/preservation activities that it undergoes;
- OreChem (Lagoze, 2009) that is used to model the chemical compounds, chemical reactions and experiments;
- OAI-ORE (Lagoze et al., 2008) that models digital objects as a bound of aggregations of Web resources.

The OPPRA ontology is formalised using OWL-DL (Bechhofer et al., 2004) that provides maximum expressiveness without losing computational completeness. A wide range of logical, and yet mature, expressions are offered by OWL-DL to achieve this goal. Examples include: 1) Boolean combinations of class expressions such as *union* and *intersection* to integrate diverse vocabularies for describing physical and digital provenance; 2) disjointness and equivalence class axioms; and 3) arbitrary cardinality restrictions.

The structure of the OPPRA ontology is developed based on the data access, acquisition and curation challenges addressed for materials informatics emerging domains (Billinge et al., 2006, Hunt, 2006). Critical requirements include: fully qualified URIs for all classes; using the *rdfs:subClassOf* construct for taxonomical relations; providing a notion of the *rdfs:label* property for human-readable descriptions; using synonyms (e.g., *oppra:hasSynonym*) for alternative definitions; defining OWL annotation properties such as *oppra:id* and *oppra:url* for references to external entities; and describing the ontology using the Dublin Core (Sugimoto et al., 2002) and its defined properties (e.g., *dc:title*, *dc:creator*, *dc:modified* and *dc:publisher*).

The OPPRA ontology is curated manually using the Stanford Protégé-OWL 4.1 (protégé, 1997). A set of OWL 2 rules for art/paint conservation are developed and applied using the OWL 2 RL (Motik et al., 2012) profile (implemented in OWLIM OpenRDF Sesame triple store (Bishop et al., 2012)) to infer new implicit relationships and knowledge from explicit data (Section 4.3.2). Inferencing is applicable to a number of aspects of art history and materials science including the relationships between:

- Condition changes and condition states of materials;
- Characterisation techniques and instruments use, calibrations and data outputs;
- Creation of materials/artefacts and their corresponding actors, artistic techniques, periods and locations;
- Physical and digital provenance of artefacts and temporal/spatial representations of data (e.g., timelines and maps).

## 4.3. Results and Discussion

This section details the classes defined by the OPPRA ontology and the class axioms and relations that have been introduced in order to accurately model the existing knowledge in paint materials and the art preservation domain. It also discusses the availability of the ontology and its envisioned revision and extension cycle.

### 4.3.1. OPPRA-specific Ontologies and Classes

The following illustrations provide details on the classes and relations existing in the OPPRA ontology. Terms with blue colours represent classes that exist in external ontologies, and the green colours represent class extensions in the OPPRA ontology. The descriptions include:

- Figure 4.1 illustrates how the CIDOC-CRM ontology can be applied to document the condition assessment and cleaning of the painting *Epiphany*;
- Figure 4.2 represents the class *Painting* with its neighbour classes;
- Figure 4.3 illustrates how the CIDOC-CRM, the OreChem ontologies and the OPPRA extensions (developed to describe paint-specific information) are combined and linked;
- Figure 4.4 represents the various condition changes that a painting may undergo (e.g., darkening, fracture, flaking, fading). The high-level concept "*ConditionChange"* is also described by OPPRA's high-level (CIDOC-CRM low-level) *E3.ConsditinState* which is a super-class of fine-grained concepts (e.g., darkened, fractured, flaked, faded) that describe their corresponding condition changes;
- Other controlled (e.g., Production, Acquisition, Assessment, Treatment, Experiment and Characterisation) and uncontrolled (Aging, Infestation, Humidity, Temperature, Radiation and Chemical Reaction) mechanisms concerning paintings are also shown in Figure 4.5;
- Figure 4.6 shows how knowledge on art conservation can be captured from different resources (e.g., condition reports, characterisation data, databases, provenance and publications) and stored using the *oai-ore:Aggregation* concepts.

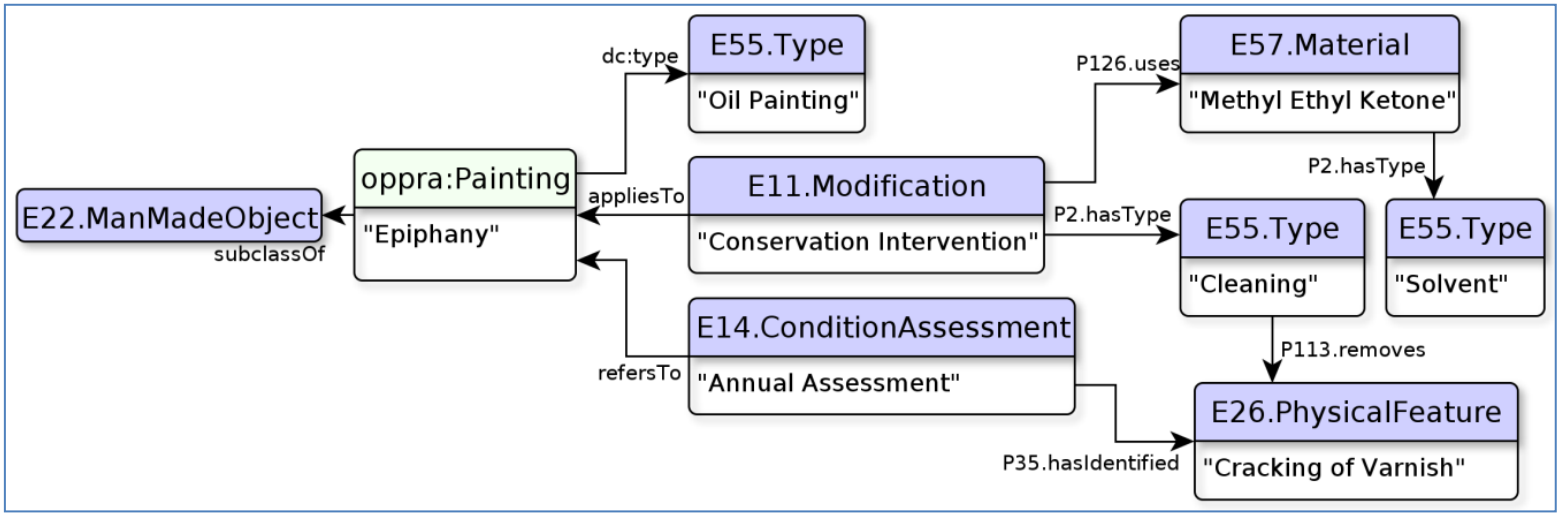



**Figure 4.1: Application of the CIDOC-CRM to painting conservation**

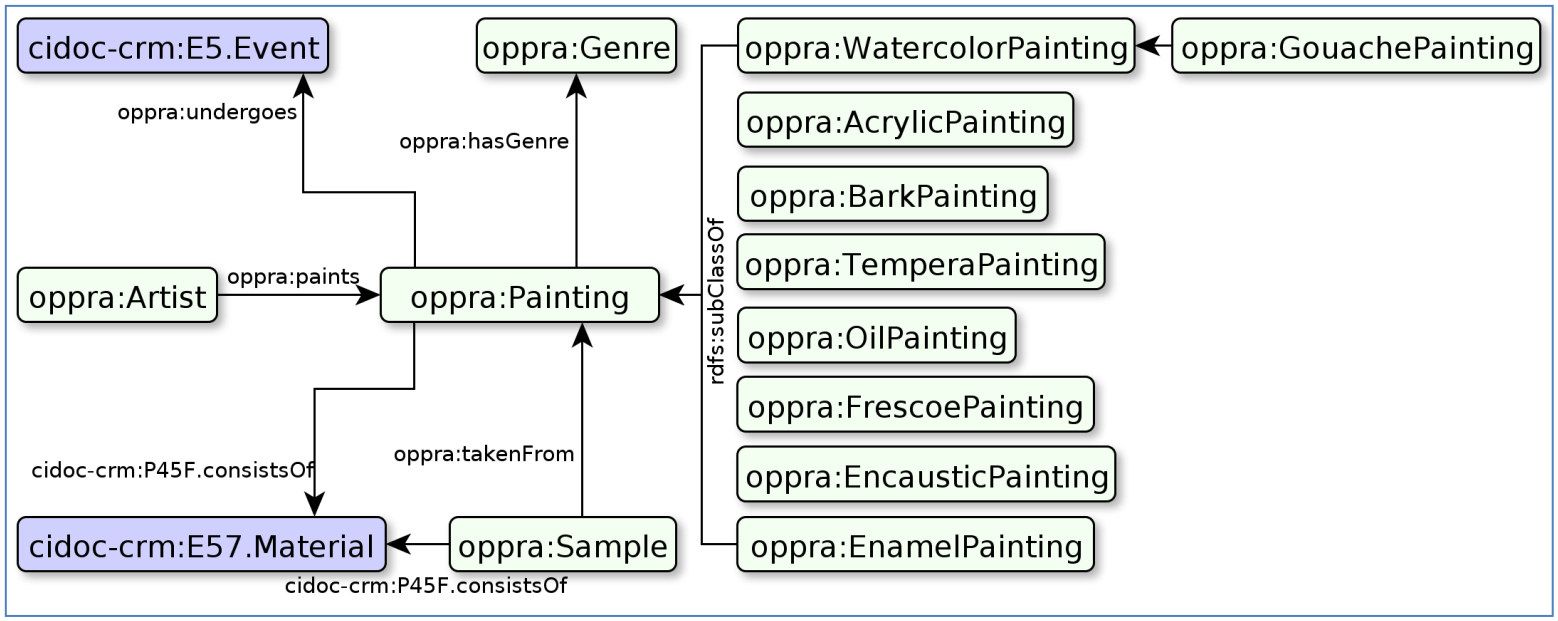



**Figure 4.2: *oppra:Painting* with its neighbour classes**

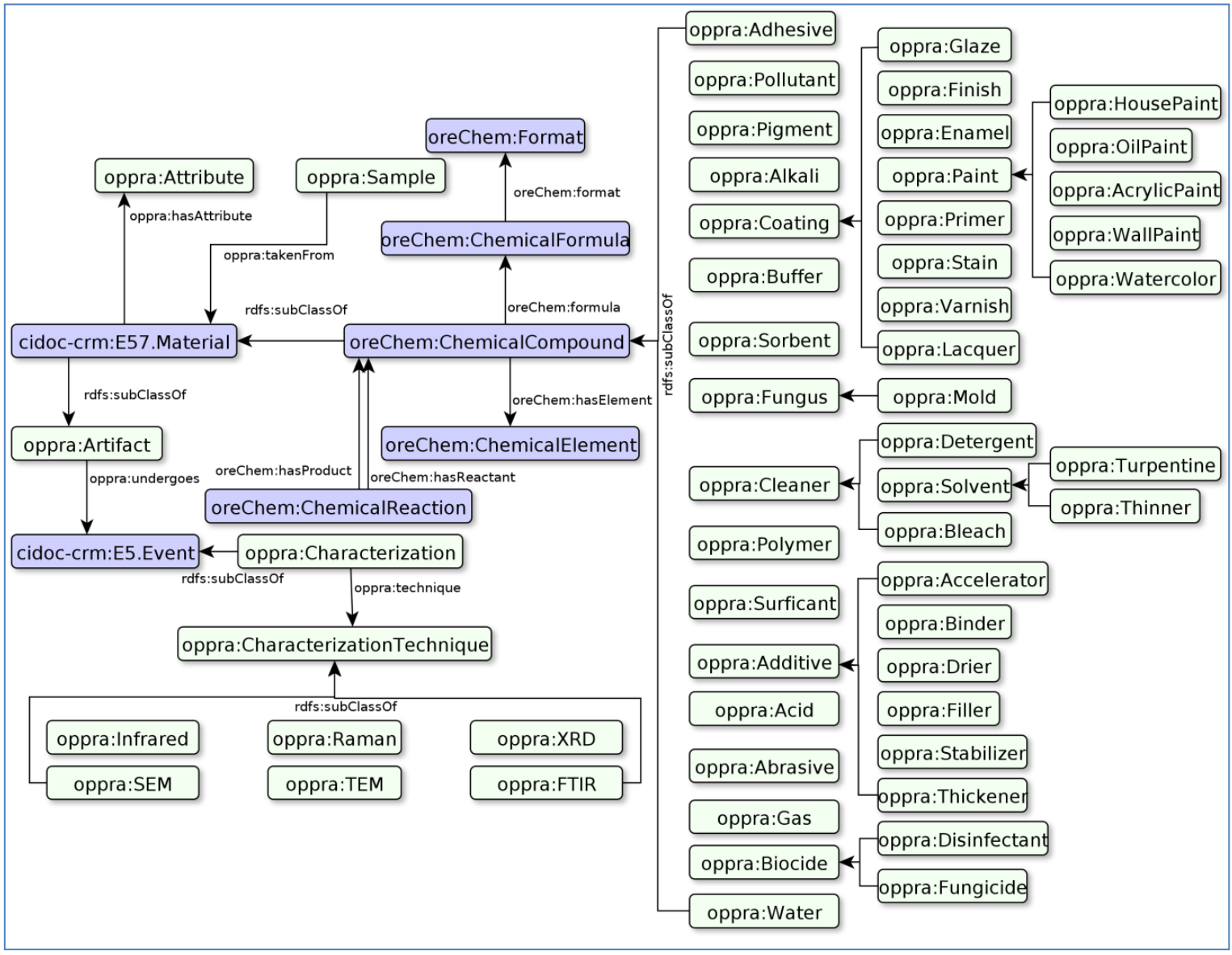



**Figure 4.3: OPPRA Extensions to CIDOC-CRM and OreChem**

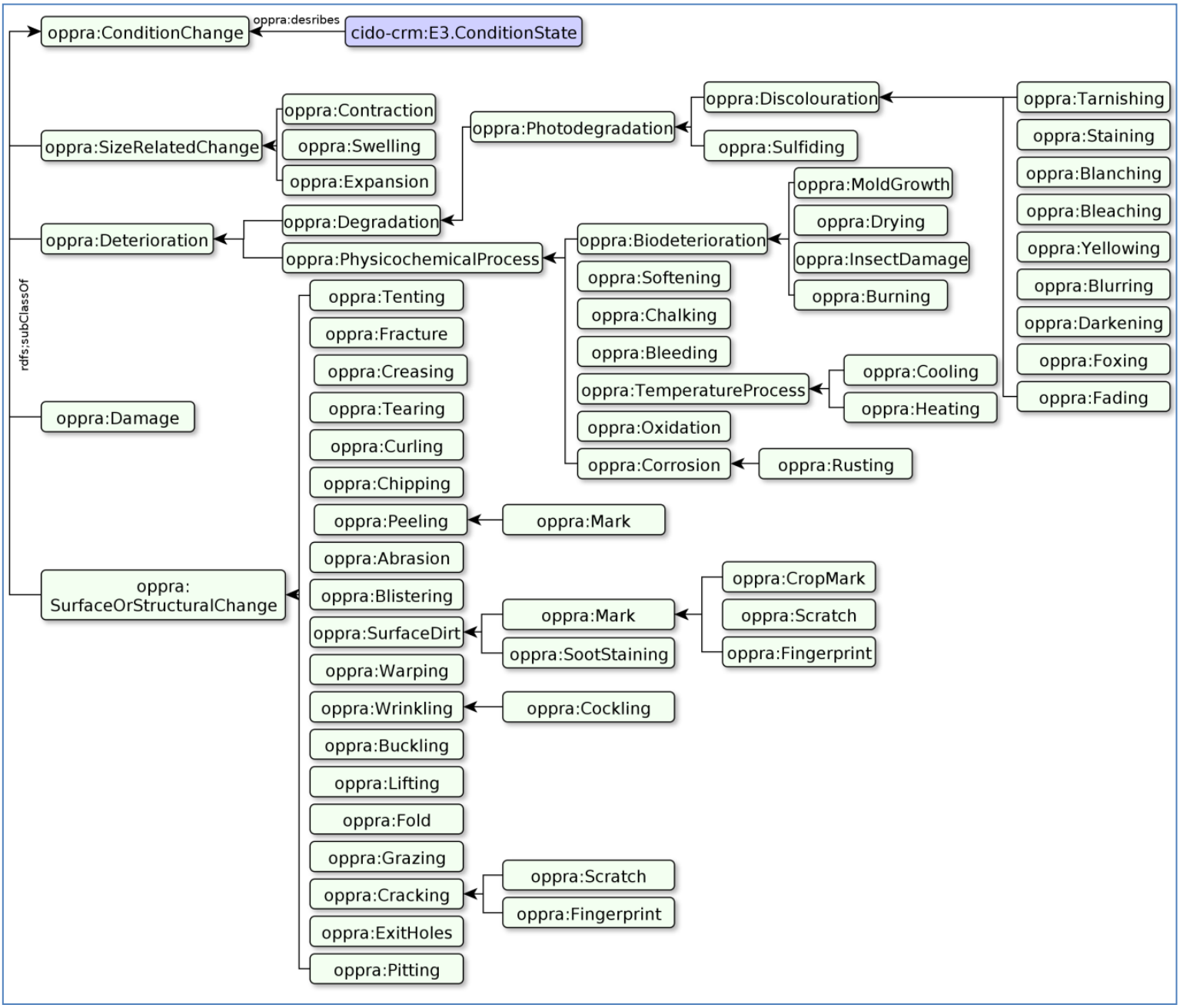



**Figure 4.4: Key concepts of *oppra:ConditionChange***

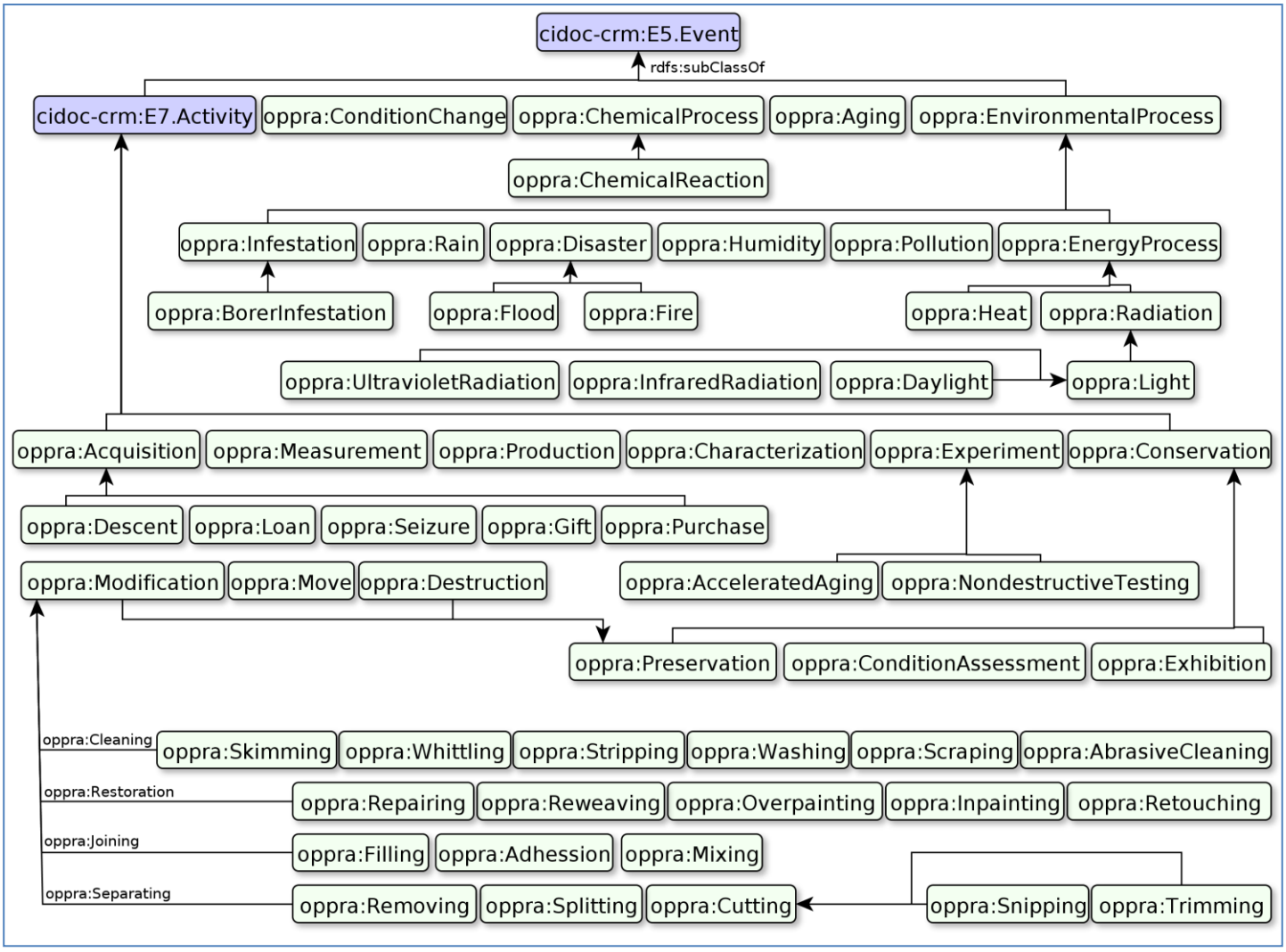



**Figure 4.5: Controlled and uncontrolled mechanisms concerning paintings**

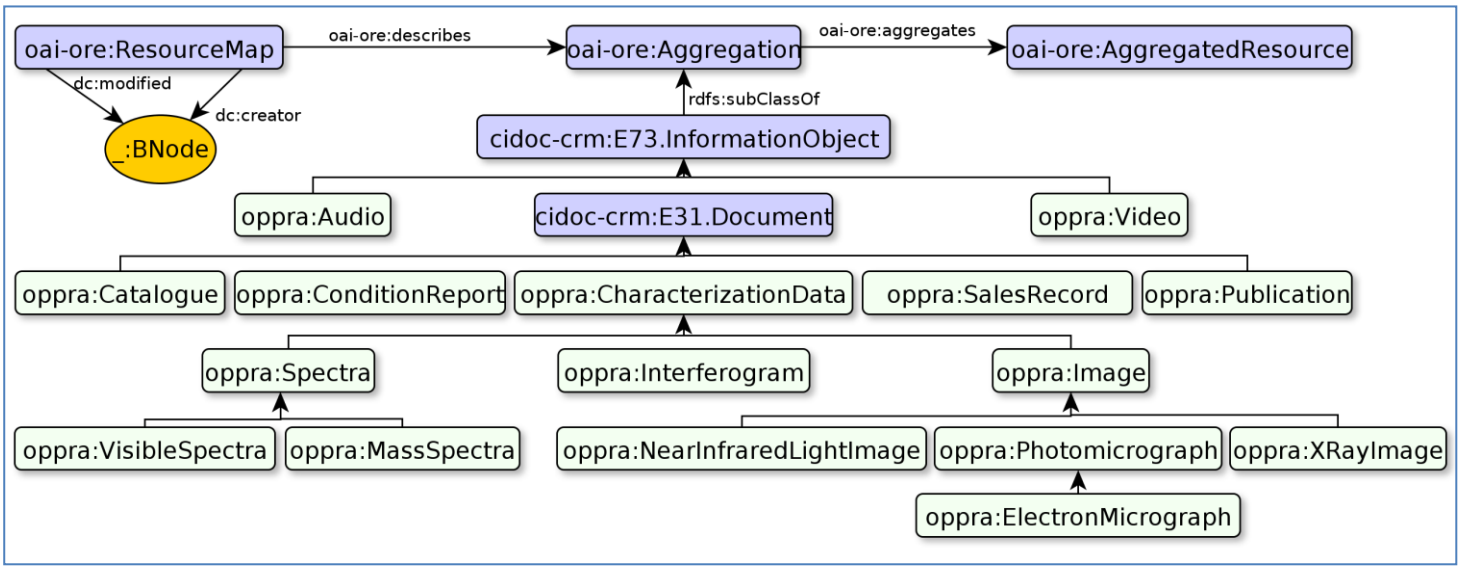



**Figure 4.6: Representation of aggregated resources on art conservation using OAI-ORE model**

### 4.3.2. Justification of CIDOC-CRM Sub-Classing Approach

An alternative approach for extending CIDOC-CRM would be to use the approach proposed by (Crofts et al., 2010) which involves using cidoc_crm:E55.Type to link to external thesauri such as the AAT, rather than creating new sub-classes of CIDOc-CRM classes (e.g., cidoc_crm:E28.Conceptual_Object *cidoc_crm.P2_has_type* cidoc_crm:E55.Type). Such an approach was adopted in several papers for automatic mapping of archaeological datasets to the CIDOC-CRM (Binding et al., 2008, Eide et al., 2008, Goodall et al., 2004, Hyvönen et al., 2009, Martinez and Isaksen, 2010, Theodoridou et al., 2010). While this is a valid approach that ensures rigorous scholarly or scientific process (Crofts et al., 2010), it would however result in difficulties associated with inferencing. For example, without sub-classing, the following aspects would have to be hard-coded into the inferencing engine – based on possible keywords used in the triples describing the domain:

- Condition changes and condition states of materials (e.g., darkened describes darkening);
- Instruments and characterisation techniques (e.g., 'Scanning Electron Microscope' wasUsedBy 'Scanning Electron Microscopy' activity);
- Painting activities and the corresponding actors ('Monastery' painting activity wasPerformedBy 'Ian Fairweather').

Furthermore, the CIDOC-CRM documentation (Crofts et al., 2010) states that:

> Users may decide to implement a concept either as a subclass extending the CRM class system or as an instance of E55 Type. A new subclass should only be created in case the concept is sufficiently stable and associated with additional explicitly modelled properties specific to it.

In this thesis, classes that are re-used from other thesauri to extend the CIDOC-CRM classes are very stable and widely adopted. The class *oppra:Artist* for example is defined by the Getty AAT (Getty, 2010b) – a well-known thesauri used within museums worldwide. Additional explicitly modelled properties and inferencing rules were also in the OPPRA extension to support a finer level of granularity for the art/paint conservation domain. For example, the following properties are associated with the class *oppra:Artist*:

- Artist *painted* Painting
- Painting *paintedBy* Artist
- Artist *hasArtisticTechnique* ArtisticTechnique
- ArtisticTechnique *artisticTechniqueOf* Artist
- Artist *performedPaintingProcess* PaintingProcess
- PaintingProcess *performedByArtist* Artist
- Artist *performedPaintingProcess* PaintingProcess AND PaintingProcess *usedArtisticTechnique* ArtisticTechnique → Artist *hasArtisticTechnique* ArtisticTechnique
- ArtisticTechnique *artisticTechniqueUsedBy* PaintingProcess AND PaintingProcess *performedByArtist* Artist → ArtisticTechnique *artisticTechniqueOf* Artist
- Artist *performedPaintingProcess* PaintingProcess AND PaintingProcess *producedPainting* Painting → Artist *painted* Painting
- Painting *wasPAintedBy* PaintingProcess AND PaintingProcess *performedByArtist* Artist → Painting *paintedBy* Artist

This level of granularity and inferencing would be difficult to achieve by linking external controlled vocabularies to *cidoc_crm::E55.Type* to create new instances of E55 Type.

### 4.3.3. Class Axioms and Relationships

Figure 4.7 lists the main relations defined within the OPPRA ontology. The *occuredInThePresenceOf* is a CIDOC-CRM relationship that links events (e.g., conservation, characterisation, painting, manufacturing, and moving) to their corresponding entities (e.g., artefact, substance, painting, and material). Examples of sub-properties of the *occuredInThePresenceOf* relation include: *assessed* that links condition assessments to paintings; *destroyed* that links destruction processes to artefacts; *hasTreated* that links treatment activities to artefacts; and *transferredTitleFrom* and *transferredTitleTo* that link acquisition activities to actors (e.g., sellers, and buyers).

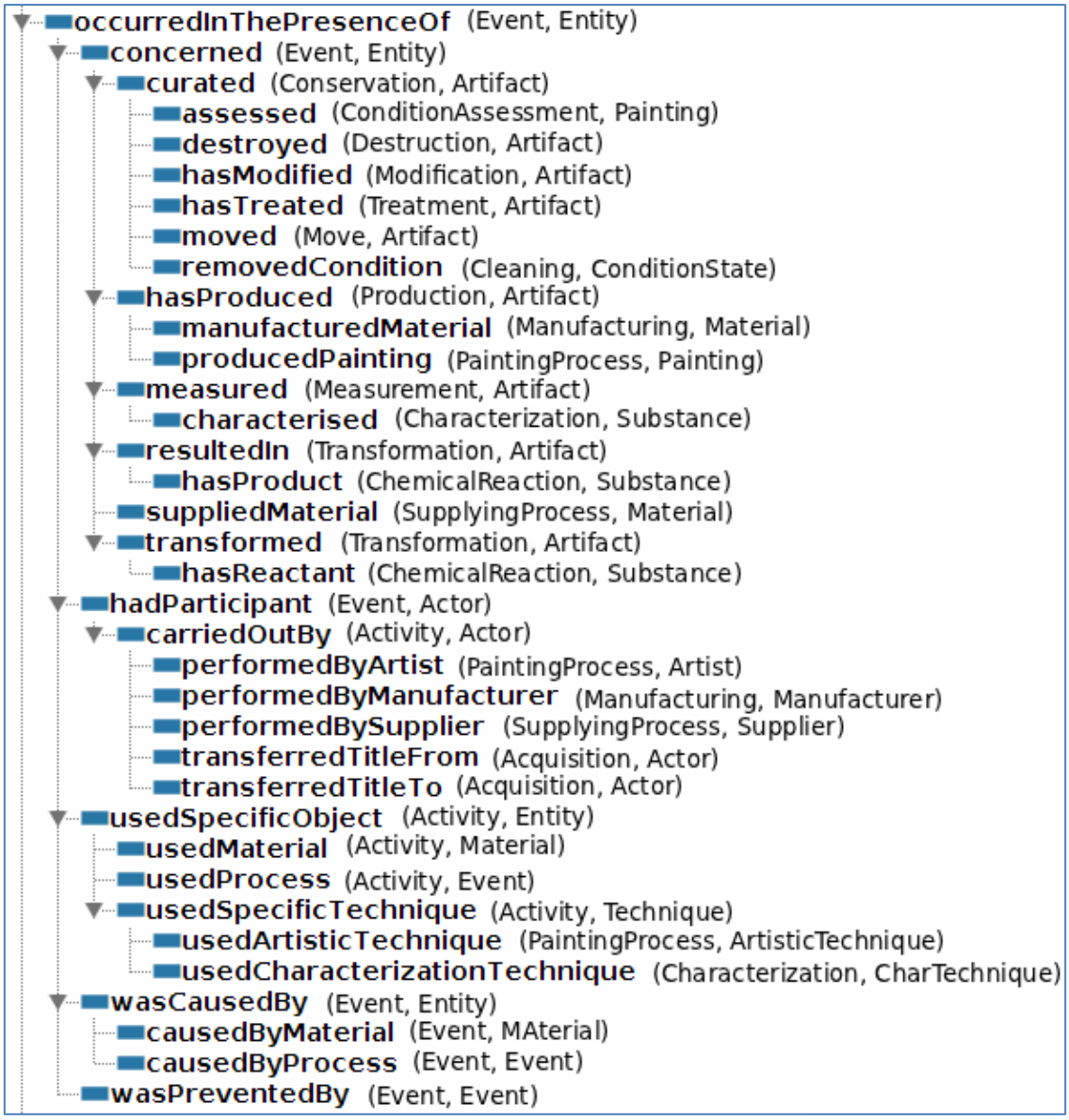



**Figure 4.7: Relations defined in the OPPRA ontology (a)**

Other top-level relationships (Figure 4.8) in the OPPRA ontology include: *assigned* that links activities to attributes (e.g.,*hasIdentified*, *observedPhysicalAttribute,* and *observedChemicalAttribute*); and *consistsOf* that generally links entities with other entities (e.g., conditions consisting of other conditions – *Yellowing*/*Cracking*, places within regions of other places – *Australia*/*Brisbane*, inscriptions within paintings and materials within artefacts – *Painting*/*Support*).

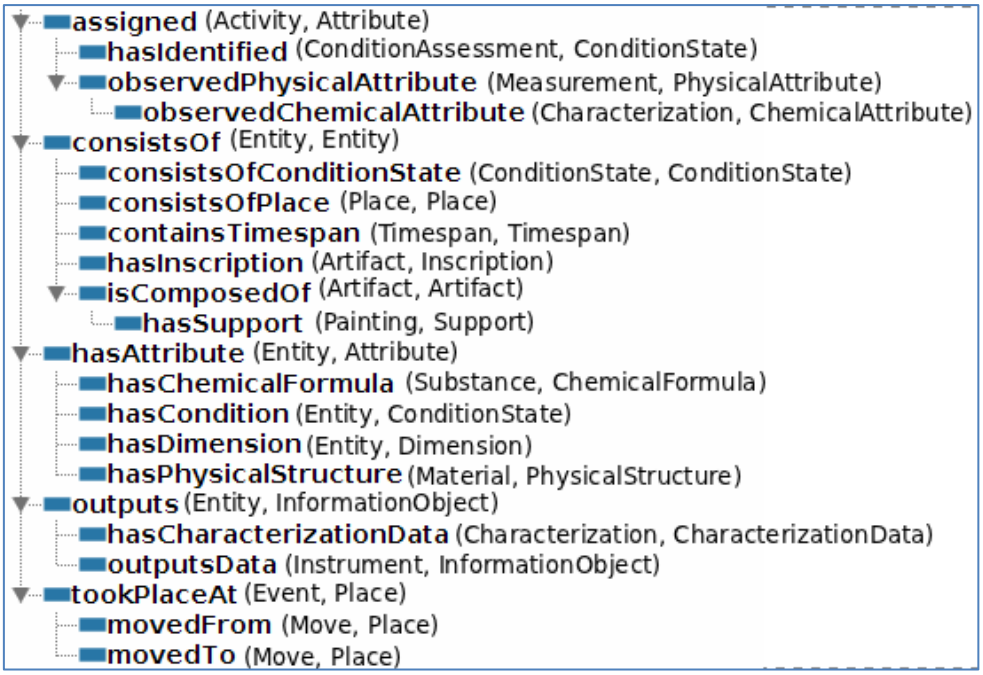



**Figure 4.8: Relations defined in the OPPRA ontology (b)**

A major aim of the OPPRA ontology is to underpin a community-driven knowledge curation platform that enables collaborative decision-making and knowledge exchange among conservators and materials scientists. In order to support knowledge capture (e.g., microscopic data, hidden knowledge within textual documents and art history, and conservation and materials science data published in external databases) as well as the decision-making processes (e.g., searching across multiple data sources, and correlating and reasoning over search results), the semantics of the emerging knowledge discoveries were encoded in class axioms and restrictions. Furthermore, to reflect the current domain knowledge about each specific activity (e.g., treatment, experiment, and characterisation) accurately, these class axioms are specialised at the lower levels of the OPPRA concept with more specific details. The class and relation axioms are implemented in OPPRA using the OWL 2 RL profile (Motik et al., 2012).

The following list provides examples of the class and relation axioms within the OPPRA ontology. In these examples, the rdfs, owl, and rdf properties are re-used within the class/relation axioms to achieve the required constrain (e.g., the explicit description of the class/class relationship, and the inferencing rule). Due to its size and complexity, it is not possible to list the full class and relation axioms in this

chapter; however, the complete OPPRA ontology is provided in (20thcpaint, 2010a), and http://www.20thcpaint.org/oppra.owl.

- ***rdfs:subClassOf***: this class axiom defines the transitive hierarchy concept between classes in the ontology.
  - Examples: (Activity, Event), (Substance, Material), (Instrument, Tool), and (Pigment, Colorant);
  - A SPARQL query example that returns all types of paintings (oil painting, watercolours, enamel paintings, acrylic paintings etc.) that were painted by the artist "*Sidney Nolan*" is: `select distinct ?painting where {?painting rdf:type oppra:Painting . ?painting oppra:paintedBy oppra:SidneyNolan}`
- ***owl:disjointWith***: this class axiom defines the disjointness restriction between sibling concepts in the OPPRA ontology – where the class extensions of the two disjoint class descriptions involved have no individuals in common.
  - Examples: (Organization, Person), (Activity, ConditionChange), (ArtisticTechnique, CharacterizationTechnique);
  - A consistency reasoner (e.g., Pellet, FaCT++ or HermiT) would throw an inconsistency exception if the following hypothetical SPARQL update statement is added: `insert data {oppra:RipolinHarmonie a oppra:AcrylicPaint, oppra:WaterBasePaint .}`
- ***rdfs:subPropertyOf***: this relation axiom defines the transitive hierarchy concept between relationships (e.g.,*owl:ObjectProperty* and *owl:DatatypeProperty*) in the ontology.
  - Examples: (hasTreated, concerned), (hasSupport, isComposedOf), and (moved, curated);
- ***owl:inverseOf***: this relation axiom defines the bi-directional relation between the *rdfs:domain* and *rdfs:range* properties.
  - Examples: (depicts, isDepictedBy), (describes, isDescribedBy), (consistsOf, formsPartOf), and (made, madeBy);
  - A SPARQL query example that returns the statements (*oppra:SidneyNolan oppra:painted ?painting* union *?painting oppra:paintedBy oppra:SidneyNolan*) is: `select distinct ?painting where{?painting oppra:paintedBy oppra:SidneyNolan}`

- ***owl:TransitiveProperty***: this relation axiom defines the transitive inferencing of concepts in the OPPRA ontology [e.g., *r(x, y) . r(y, z)* → *r(x, z)*].
  - Examples: *rdfs:subClassOf, rdf:type, oai-ore:aggregates, oppra:takenFrom, oppra:hasMaterial*;
  - If a sentence includes the triple/statement oppra:hasDescription (oppra:Experiment, "light bleaching"), then a SPARQL query that returns all publications that include light bleaching experiments is: `select distinct ?publication where{?publication oai-ore:aggregates ?sent . graph ?sent{oppra:Experiment oppra:hasDescription "light bleaching"} }`
- ***rdfs:domain***: this relation axiom defines the class of the subject in a given triple.
  - Examples: outputs(Characterisation), undergoes(Artifact), and usesMaterial(Activity);
- ***rdfs:range***: this relation axiom defines the class of the object in a given triple.
  - Examples: outputs(CharacterisationData), undergoes(Event), and usesMaterial(Material);
- ***owl:propertyChainAxiom***: this relation axiom connects all individuals that are linked by a chain of two or more object properties. The following examples provide specific details of *owl:propertyChainAxiom*:
  - oppra:paintedBy(oppra:Painting, oppra:Artist) ←→ oppra:carriedOutBy(oppra:PaintingProcess, oppra:Artist) . oppra:concerned(oppra:PantingProcess, oppra:Painting);
  - oppra:hasCondition(oppra:Painting, oppra:ConditionState) . oppra:describes(oppra:ConditionState, oppra:ConditionChange) ←→ oppra:undergoes(Painting, ConditionChange). This also applies to fine-grained types of condition changes and condition states, such as (oppra:*Yellowing*, oppra:*Yellowed*), (oppra:*Bleaching*, oppra:*Bleached*), (oppra:*Blistering*, oppra:*Blistered*), (oppra:*Wrinkling*, oppra:*Wrinkled*);
  - oppra:removesCondition(oppra:Cleaning, oppra:ConditionState) ←→ oppra:concerns(oppra:Cleaning, oppra:Painting). oppra:hasCondition(oppra:Painting, oppra:ConditionState);
  - oppra:characterisationTechnique(oppra:Activity, oppra:CharacterisationTechnique) ←→ oppra:uses(oppra:Activity,

oppra:Instrument) . oppra:usedInTechnique(oppra:Instrument, oppra:CharacterisationTechnique).

### 4.3.4. Availability

Table 4.1 summarises the main characteristics of the OPPRA ontology. The current release of the ontology has the version number 1.0, and the namespace of the ontology is http://www.20thcpaint.org/oppra.owl.

**Table 4.1: The OPPRA ontology fact sheet**

| Name | OPPRA |
|---|---|
| Namespace | http://www.20thcpaint.org/oppra.owl |
| Prefix | oppra |
| Format | OWL-DL |
| Number of classes | 2325 |
| Number of relations | 169 (Object properties)<br>12 (Data properties)<br>7386 (Annotation properties) |
| Dependencies | owl, rdf, rdfs, xsd, dc, cidoc-crm, oreChem, oai-ore |
| Number of axioms | 2325 (rdfs:subClassOf)<br>920 (owl:disjointWith)<br>169 (rdfs:subPropertyOf)<br>169 (owl:inverseOf)<br>5 (owl:TransitiveProperty)<br>181 (rdfs:domain)<br>181 (rdfs:range)<br>32 (InverseObjectProperties) |

Figure 4.9 also shows a screenshot of the Web interface that allows readers to browse the ontology (20thcpaint, 2010a). This interface was originally developed to enable collaborators to comment on and provide feedback to the draft ontology – which evolved through an iterative cyclical process. The Web version was dynamically produced using Protégé's OWLDoc plug-in that exports views of the ontology as HTML. The figure shows the class hierarchy associated with the class *SyntheticResinPaint* – it is a sub-class of *Paint* and a super-class of *AcrylicPaint*, *VinylPaint* and *PolymerPaint*.

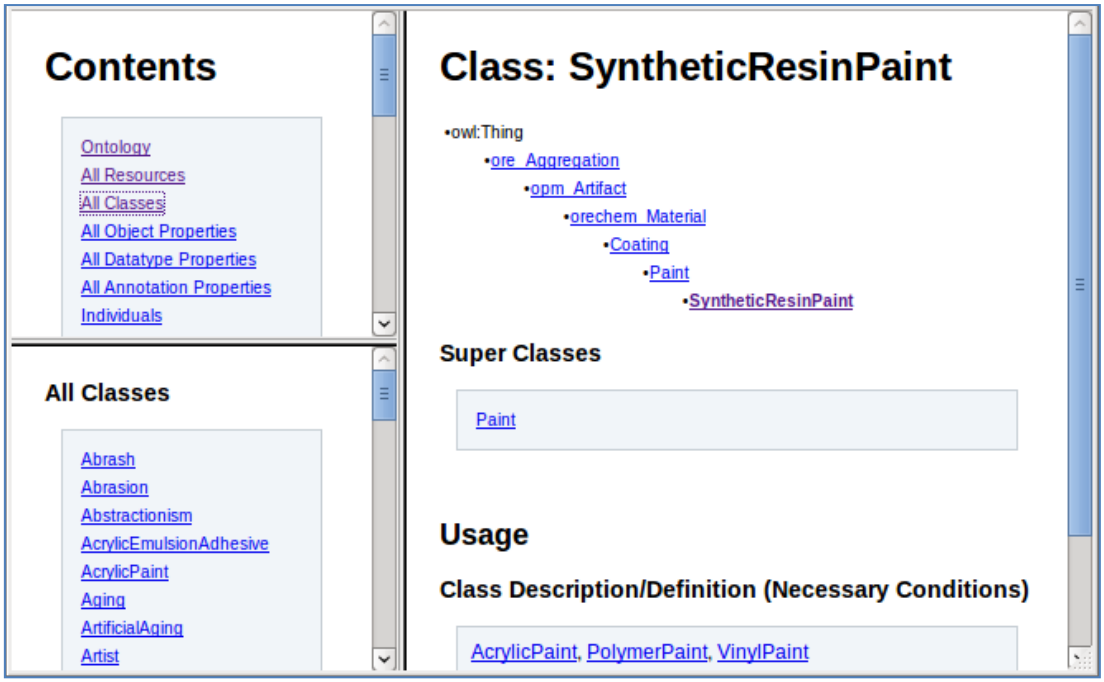



**Figure 4.9: Web interface of the OPPRA ontology**

## 4.4. Comparison and Evaluation Criteria

### 4.4.1. Comparison to Related Ontological Resources

Among the specific ontologies mentioned above (Section 4.3.1), the actual painting and preservation of art knowledge (representing the core of the OPPRA ontology) is covered only superficially in other ontologies and vocabularies. For example, ontologies such as CIDOC-CRM, OreChem, and OAI-ORE denote high-level concepts that correspond to their particular domain (e.g., provenance of painting and conservation activities, chemical compounds and reactions, as well as digital aggregations of Web resources).

In addition, controlled vocabularies that provide information on art history and materials science concepts do not express paintings' provenance and preservation techniques in a meaningful way – in a multi-dimensional concept/relation/axiom representation. Examples of such vocabularies include: AAT (Getty, 2010b) that structures vocabularies to improve access to information about art, architecture and material culture; CAMEO (MFA-Boston, 1997) that provides a material database containing chemical, physical, visual and analytical information on historic and contemporary materials used in the production and conservation of artistic, architectural, archaeological and anthropological materials; and the AICCM Visual

Glossary (AICCM, 1999) that provides artefact collectors and curators a way to explore and describe the condition states of art objects.

The aim of the OPPRA ontology is to develop, curate and share controlled vocabularies to support the evolving knowledge in the art history and conservation science domains. The goal of the OPPRA ontology is to bridge the gap between the physical (e.g., deterioration, condition assessment, exhibition, movement, and treatment) and digital (e.g., paint material, characterisation, physical/chemical structure, and degradation mechanism) provenance of paintings in order to build a comprehensive body of knowledge from existing and emerging preservation techniques.

The added value of the OPPRA ontology stands in its comprehensive classification and accurate description (via class and relation axioms) of the physical and digital provenance of paintings. The other ontologies, in particular CIDOC-CRM, OreChem and OAI-ORE, are regarded as effective complementary and important resources to be cross-referenced and re-used (to avoid redundancy) to describe the provenance of paintings.

To date, the integrity of the OPPRA ontology has been ensured by three teams of conservators and materials scientists within the 20$^{th}$ Century in Paint project. The initial testing of its applicability is reported in this thesis (Chapters 6-9) and will be further evidenced by the extent of its changes over time and the future growth of the 20$^{th}$ Century in Paint knowledge-base and its associated community of users.

However, two limitations were identified, and deserve further investigation to improve the OPPRA ontology over time:

- The OPPRA ontology is currently limited with regard to certain specific high-level concepts that are significant within the art/paint conservation domain such as time and temporal relations (e.g., Time ontology (Hobbs and Pan, 2006)), or place and spatial relations (e.g., Geospatial ontology (Lieberman et al., 2007)). It is believed that the OPPRA ontology is able to incorporate such additional ontologies through extensions, in the same way it incorporates the OreChem ontology.

- Currently there is no interface that enables the art/paint conservation community to interactively and collaboratively edit/refine the OPPRA ontology. Provision of an online easy-to-use collaborative editing interface, accessible to authenticated experts, would be the quickest and most efficient way to improve the ontology over time. Furthermore, investigating the best ontology library for publishing the OPPRA ontology to the Semantic Web and exposing the ontology to the art/paint conservation community is worth pursuing.

### 4.4.2. Quality of the OPPRA Ontology

The quality of the OPPRA ontology is assessed based on the following five criteria (Gruber, 1995): clarity, coherence, extensibility, minimal encoding bias, and minimal ontological commitment. Satisfactory results are achieved. This methodology has been successfully used previously to evaluate ontologies within many informatics fields (Abu et al., 2013, Cheung et al., 2008, Sidhu et al., 2007).

*Clarity*

Definitions within an ontology need to be stated in such a way that the number of possible interpretations of a concept would be restricted. This will contribute to the effectiveness of communication between agents. In the design of the OPPRA ontology, it is stated that for each concept *c* with property *p*, the pair *(c, p)* exactly specifies a unique pair. During the design of the ontology this rule is enforced, and the uniqueness of the definition of concepts is guaranteed. Clarity of the OPPRA ontology is also checked by running the tests listed below and making sure all of them return true:

- **No cardinality restriction on transitive properties, and transitive properties cannot be functional**. Data in the art/paint conservation domain is evolving over time whereby a new data type may need to be inserted into the ontology at any time. Thus, for transitive properties, cardinality restrictions are not assigned. In addition, these transitive properties cannot be functional because they relate to more than one instance via inferencing.
- **No classes or properties in enumerations**. As seen throughout the OPPRA ontology description (Chapter 4), and its OWL 2 representation (20thcpaint, 2010a), it is clearly shown that there are no classes or properties in enumeration.

- **No import of system ontologies**. Even though the CIDOC-CRM, OreChem, and OAI-ORE ontologies are re-used, and terms from AAT, CAMEO, and AICCM visual glossary are adopted (as discussed in Chapter 4), the OPPRA ontology is designed to satisfy the specific requirements of the art/paint conservation community. Thus, other system ontologies are not directly imported. The terms extracted from the external glossaries (e.g., AAT, CAMEO, and AICCM visual glossary) were entered manually under their appropriate classes in the OPPRA ontology. The class/subclass locations of these terms in the ontology were validated by reviews undertaken by the art conservators and materials scientists working on the 20$^{th}$ Century in Paint project.
- **No meta-class, and no subClasseOf RDF classes**. As also seen throughout the OPPRA ontology description (this chapter), and its OWL 2 representation (20thcpaint 2010a), there is no meta-class, and no subClasseOf RDF classes.
- **No super or sub-properties of annotation properties**. There are no super or sub-properties of annotation properties used in OPPRA, because the built-in *Annotation* property in Protégé (protégé, 1997) is used throughout the class, object property, data property, and instance descriptions.

Furthermore, OPPRA possesses clarity because its vocabulary is sourced from peer-reviewed ontologies (CIDOC-CRM, OreChem, and OAI-ORE) and existing standardised taxonomies (AAT, CAMEO, and AICCM visual glossary).

*Coherence*

The definitions of concepts given in the ontology should be consistent. Only inferences consistent with existing definitions should be allowed. The formal part of the OPPRA ontology is checked by running the consistency tests listed below and ensuring that, for these tests, all return true:

- **Domains and ranges (of properties) should not be empty, and should not contain redundant classes**. As seen in the OWL 2 representation of OPPRA (20thcpaint 2010a), all properties are assigned only one domain, and range. Thus, the domain and range of properties in OPPRA are not empty, and do not contain redundant classes.
- **The relation between the super/sub-level property, and its domain/range super/sub-level class must be consistent as follows**:

- Domain and range of a sub-property can only narrow its super-property;
- Inverse of sub-property must be sub-property of inverse of super-property;
- Inverse of top-level property must be top-level property;
- Inverse of functional property must be 'inverse functional';
- Inverse of 'inverse functional' must be functional property;
- Inverse of symmetric property must be symmetric property;
- Inverse of transitive property must be transitive property;
- Inverse property must have matching range and domain.

Furthermore, OPPRA has no coherency issues because there are no concepts derived via inferencing. The ontology is created by hand (based on the art/paint conservation community's requirements) using Protégé. The ontology-based tools (Experimental Data Capture, *text2triples*, data extracted from external databases, and inferencing) only add RDF instances to the knowledge-base (i.e., no classes, properties, and inferencing rules are added at any stage of the data capture process).

*Extensibility*

It should be possible to extend the OPPRA ontology without altering the existing definitions. Easy ontology extension is quite an important feature as new knowledge emerges each day, and may need to be added to an already existing ontology. To make OPPRA extendable, the design (represented by OWL 2) consists of a hierarchical classification of concepts represented as classes, from general to specific. In OPPRA, the notions of classification, reasoning, and consistency are applied by defining new concepts from defined generic concepts. The concepts derived from generic concepts are placed precisely into the class hierarchy of the OPPRA ontology, to completely represent the information defining the art/paint conservation entities. Thus, this ontology does not sanction a preference for one class (e.g., Material) only, and allows for the definition of other classes (e.g., ChemicalCompound), and a way to relate them to existing classes (e.g., isPartOf).

Furthermore, OPPRA is extensible because CIDOC-CRM, OreChem, and OAI-ORE provide proven models for integrating multi-disciplinary ontologies (high-level

provenance ontologies that provide a platform for integrating ontologies within museums, and e-science artefacts).

*Minimal Encoding Bias*

The ontology representation language should be as independent as possible from the use of the ontology. In developing the OPPRA ontology, the choice of representation language (OWL 2) keeps the encoding bias to a minimum – as the ontology is intended to be used by the art/paint conservation community (e.g., art historians, conservators, curators, scientists, and researchers).OPPRA has no encoding bias because it is free of implementation details.

*Minimal Ontological Commitment*

Ontologies should make as few claims as possible about the domain while still supporting the intended knowledge sharing. The OPPRA ontology has an ontological commitment that is as low as domain ontology, because it re-uses most of the concepts that have already been used to represent art/paint conservation data and knowledge, and proposes fewer new concepts (i.e., because existing peer-reviewed ontologies are re-used and extended on standardised vocabularies). The low ontology commitment of the OPPRA ontology makes it more extendible and re-usable as shown above. Also, if fewer new concepts need to be agreed upon by the art/paint conservation community, then this makes agreement easier.

### 4.4.3. OPPRA's Achievements and Querying Capabilities

Within the context of the 20th Century in Paint project, and as will be seen throughout Chapters 6-8, OPPRA has been successfully used to:

- Document and describe experiments conducted by conservators and scientists, and allow them to upload their data and findings, link, share and re-use this data, and make it accessible publicly – with efficient storage and user experience. This aspect is described in Chapter 6.
- Automatically extract structured data about past research and experiments from relevant publications. The ontology was used to identify and markup key entities and concepts described in the publications (NER with 93.16% accuracy), and to build a structured case study (RE with 60.02%-79.09% accuracy).This aspect is described in Chapter 7.

- Aggregate information, and reason over the information, from internal and external datasets – to answer advanced and semantically linked queries. This aspect is described in Chapter 8.
- Provide a common, machine-readable formal representation of the knowledge in the domain of art/paint preservation. This aspect is also described in Chapter 8.
- Bridge the gap between the physical and digital provenance of paintings – to build a comprehensive body of knowledge from existing and emerging art preservation techniques. This aspect is also described in Chapter 8.

## 4.5. Summary

OPPRA – described in this chapter represents the first attempt to provide a comprehensive knowledge representation to the art/paint conservation domain. It provides the means for documenting (in a machine-processable form) the artefacts, events, agents and information objects that are involved in the art/paint preservation/conservation domain.

This chapter focused on a detailed description of the existing ontologies, and the classes, properties and rules that comprise the OPPRA ontology. The results of the OPPRA evaluation are given, and will be evidenced by Chapters 6 to 8. Finally, future work for the completeness of the OPPRA ontology involves: incorporating additional ontologies (through extensions) such as time and temporal relations (e.g., Time ontology), and place and spatial relations(e.g., Geospatial ontology); provisioning of an online easy-to-use collaborative editing interface, that is accessible to authenticated experts; and investigating the best ontology library for publishing the OPPRA ontology to the Semantic Web and exposing the ontology to the art/paint conservation community.

# Chapter 5

# Designing the Technical Framework

## 5.1. Introduction

As discussed in Chapters 1-4, a major aim of this project is to enable art/paint conservators to search, query and aggregate relevant data and metadata that are highly heterogeneous and distributed across databases, scholarly publications and the Web – to find answers to the complex issues they are trying to understand and solutions to the problems they are trying to solve.

This chapter focuses on the knowledge-base and technical framework (a community-driven knowledge curation platform for the art/paint conservation domain) that underpin the tools and services developed for the 20$^{th}$ Century in Paint project. As outlined in Chapters 1 and 3, the following list summarises the requirements of the art conservators and materials scientists (as identified via the APTCAARN member workshops):

- Support for the typical art/paint conservation workflow (described in Chapter 3);
- An online repository (or repositories) to capture, describe, share and re-use the results obtained from experiments investigating the causes of paint degradation/deterioration, the optimum treatments and the best methods for prevention and preservation;

- The ability to extract structured data about past research and experiments from relevant publications or other textual documents (technical reports, Websites, Wikis) on art conservation, and store it in the knowledge-base – to enable fast, easy access to and comparison across related information;
- The ability to link, search, correlate and integrate relevant data and information (from local databases, external databases and related publications) on art materials, paints, paint deterioration mechanisms, paint characterisation data, conservation techniques, provenance, and artistic practices;
- Persistent links (URLs) to the provided resources (via publications) – to reliably discover these resources as they are moved, removed or renamed;
- The ability to protect their results through authenticated access control mechanisms until they are ready to publish/share them;
- An integrated and event-aware informatics framework/knowledge-base that will:
    - Establish a semantic Linked Open Data network – linking art history, paintings, people, artistic techniques, provenance, samples, experiments, characterisations and preservation treatments;
    - Provide answers to more sophisticated queries than traditional databases that provide silos of information about specific aspects of art conservation (e.g., Give a list of oil paintings showing cracks due to metal soap formation);
    - Provide a decision support tool to recommend the most appropriate method given a specific art conservation problem (e.g., the best solution for cleaning mould from a 20$^{th}$ century acrylic painting by Ian Fairweather);
- A mechanism and proposed standardised approach for linking experimental and/or characterisation data to publications.

The following sections describe the design options and components developed to support these requirements.

## 5.2. The 20$^{th}$ Century in Paint Platform

The goal in this thesis is to provide the conservation community (initially as a test-bed within the 20$^{th}$ Century in Paint project) with a scalable, federated, distributed data management solution – a secure Web portal and Wiki that provide different

levels of collaborative access to data, models, services and storage regarding industrial and artist paints.

A Web framework that would satisfy the requirements above is needed. For the purposes of rapid development, the following decisions were made with regard to the best technologies for the framework:

- Web interface – HTML, CSS, JavaScript, AJAX, Dojo, Apache Tomcat, JSP, Apache, PHP to ensure dynamic Web interfaces and highly responsive interactivity;
- Wiki – MediaWiki to enable informal knowledge sharing among the team members;
- Security – XACML to provide a fine-grained access control mechanism;
- Persistency/Databases – Jackrabbit to store images and spectra, and MySQL to store information about users, pages, contents, revisions, file metadata, and security information such as access rights and encryption keys;
- Knowledge (databases, publications) linking – OWL to provide standardised machine-processable metadata schemas, vocabularies and formats;
- Inferencing and searching – Sesame, OpenRDF, OWLIM to store metadata (knowledge – RDF statements) and provide the means for this knowledge to be accessed (e.g., SPARQL), inferred (e.g., OWL 2 RL) and re-used (e.g., RDF, N3, Turtle, TriX, TriG);
- Text analysis – GATE to read text (e.g., PDF, HTML, TXT, DOC), to process the text using the provided extensions (e.g., tokenisation, sentence splitting, PoS tagging, and morphological analysis), and to allow a customised pipeline to be executed (e.g., NER, ambiguities resolution, and RE);
- Machine Learning – ML (e.g., CRF, MALLET) to perform automatic NER, ambiguities resolution, and RE.

An overview of the technical architectural framework is shown in Figure 5.1. The following sub-sections provide details on each of the following principal components within the 20$^{th}$ Century in Paint project: the Web portal (public, members and Wiki areas), authentication and access control, the OPPRA ontology, inferencing and the underlying triple store, experimental data capture and workflow management

system, structured data extraction from past publications, and data linking, querying and visualisation tools.

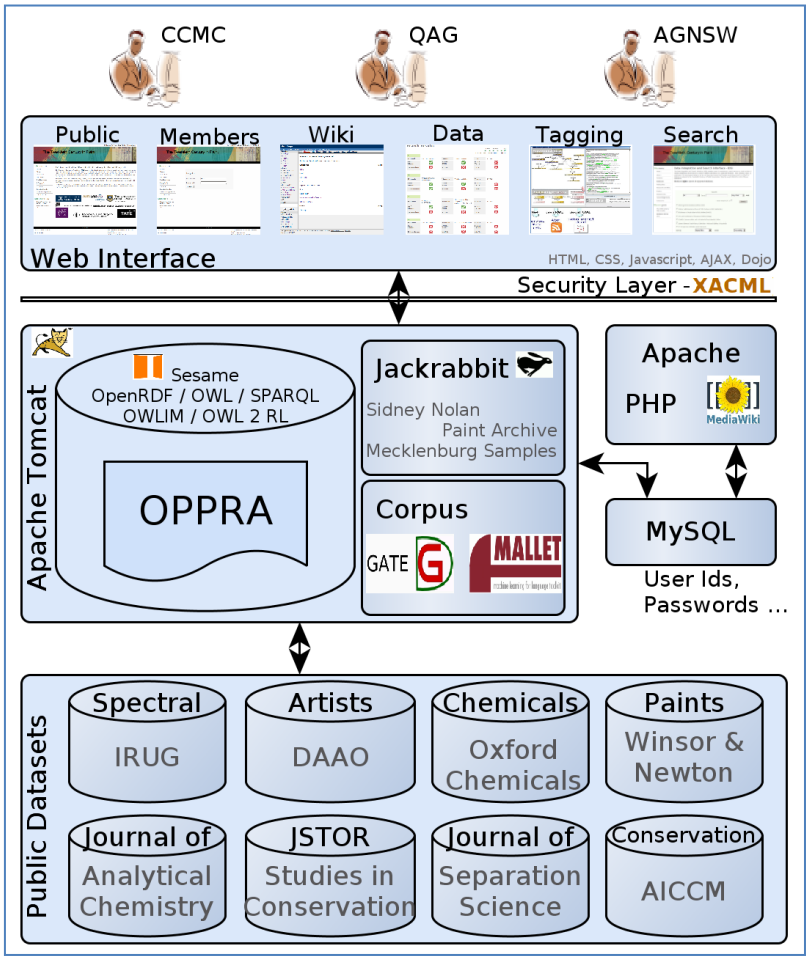



**Figure 5.1: High-level architecture of the 20th Century in Paint Web portal**

### 5.2.1. Web Portal – Public, Members and Wiki Areas

The 20th Century in Paint Web portal (20thcpaint, 2010b) comprises several Web-based interfaces that enable users (collaborating scientists and conservators) to interact with the provided services. The entry page (illustrated in Figure 5.2) is the public Website that includes links to the public and members' areas such as: about (information about the project, backgrounds, aims, approach and significance); project members (collaborating institutions and people); research activities (detailed research activities by teams 1-4); databases (an entry point to members' services – Sidney Nolan Paint Archive, Mecklenburg Samples, the OPPRA ontology browser, *text2triples*, and Integration, Search and Visualisation services); publications by the collaborators; recent and past events and activities conducted by the collaborators; related links; logos and acknowledgments; and contact information.

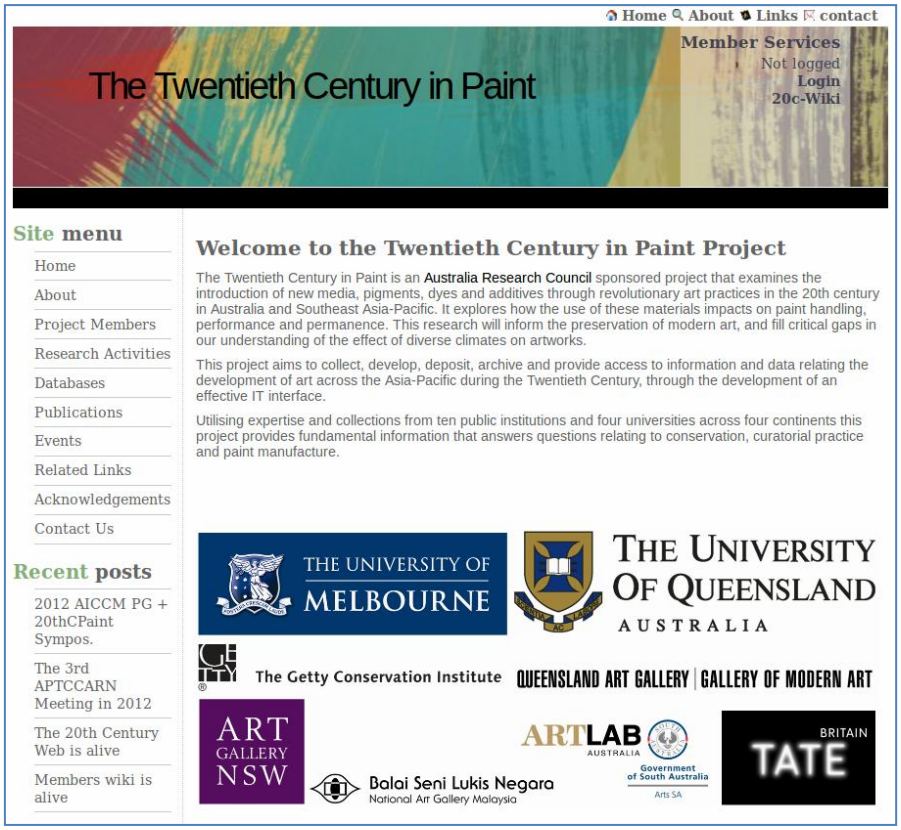



**Figure 5.2: Web portal – public and members' links (20thcpaint, 2010b)**

Other links provided in the header section are: the login page that enables members' interactions based on the security aspects (discussed below in Section 5.3.2); and the Wiki area (MediaWiki as shown in Figure 5.3) that allows collaborators to exchange information about their research activities, notifies collaborators about meetings, and provides access to the project's forms and progress reports.

### 5.2.2. Authentication and Access Control

Access control is imposed on each component of the Web portal. The granularity of the view depends on the privileges and access policies for authenticated users – enforced and defined by Jackrabbit, MediaWiki and XACML security measures.

To login and interact with the system, each conservator/scientist is prompted to enter his/her username and password by an encrypted HTML form. The login form is mapped to the Jackrabbit Login Module that authenticates users and assigns access policies on each successful login.

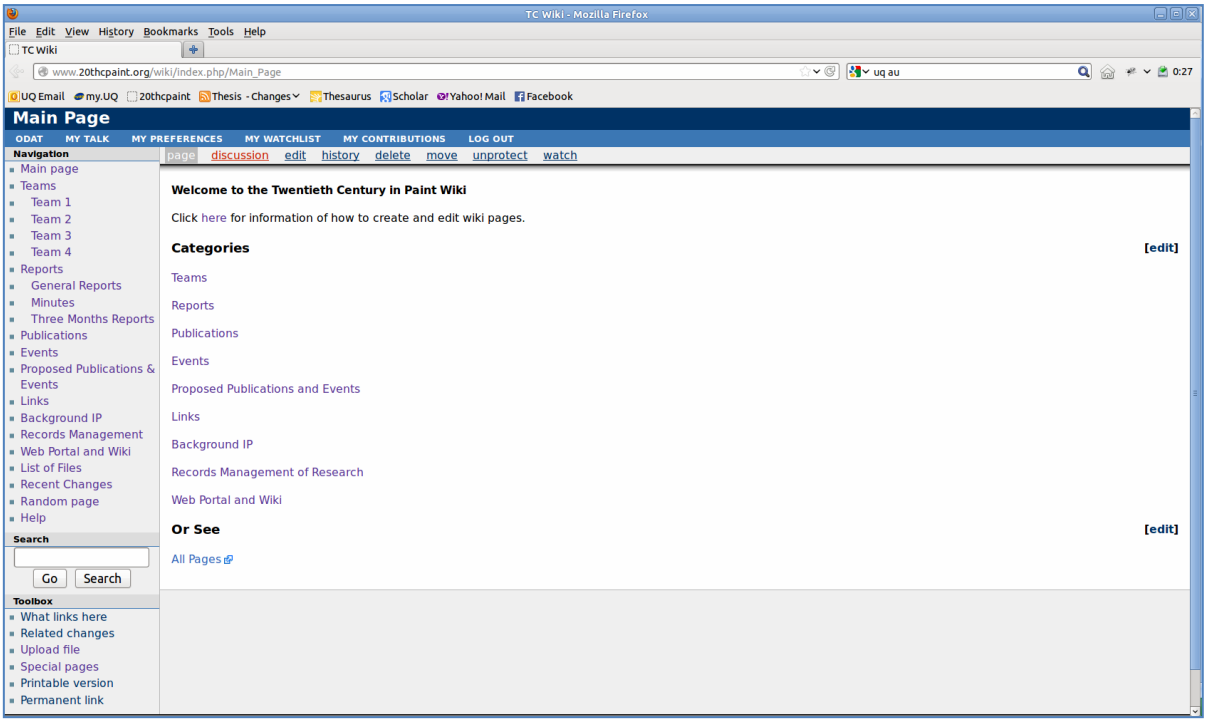



**Figure 5.3: Wiki area (MediaWiki) – members' links and information exchange**

For authentication, the Jackrabbit Login Module accesses the MediaWiki database for a list of usernames and encrypted passwords, checks the users' credentials, and returns the results back to the server to save the session of a logged/unlogged user. The decision to share the users' credentials with MediaWiki was made to unify users' login information, and to allow a single sign-in operation between the public website, members' workspaces and MediaWiki.

After a successful authentication, fine-grained access rights are enforced using the extensible Access Control Markup Language – XACML Java implementation and access control policies set in advance for each user. For example, the user "*odat*" is set as an administrator who can change access rights, reset passwords and perform administrative tasks as instructed by the project managers.

### 5.2.3. The OPPRA Ontology and Underlying Triple Store

The OPPRA ontology underpins the services provided to the 20th Century in Paint project teams. It provides a mediation component between the data storage (e.g., Jackrabbit repositories), data extraction (e.g., text analysis software - *text2triples* and data capture from external datasets), inferencing (e.g., OWL 2 RL) and end-user services (e.g., SPARQL queries and visualisation tools).

The OPPRA ontology includes inferencing rules which are implemented using OWL 2 RL (Motik et al., 2012) (to provide maximum expressiveness without losing computational power). Chapter 4 provided examples of the OWL 2 RL rules (class and relation axioms) existing in the OPPRA ontology (e.g.,*rdfs:subClassOf, owl:disjointWith, rdfs:subPropertyOf, owl:inverseOf, owl:TransitiveProperty, rdfs:domain, rdfs:range* and *owl:propertyChainAxiom*).

Chapter 4 also showed the ontology browser (20thcpaint, 2010a) developed to enable collaborators to comment on and provide feedback to the draft ontology – which evolved through an iterative cyclical process.

Once the art/paint preservation knowledge (experimental data and knowledge from publications and external databases) is captured and published, it is stored in the OpenRDF triple store (in a form that complies with OPPRA). The RDF representation of published data makes it easy for human and computer agents to perform queries across the knowledge-base (for further search, export and visualisation tasks).

### 5.2.4. Experimental Data Capture and Workflow Management

The Experimental Data Capture system enables collaborators within the 20$^{th}$ Century in Paint project to capture and record all the data generated by their activities (e.g., sampling – ids, materials, codes and locations; observations – physical and chemical attributes; characterisation data – images and associated metadata). Figure 5.4 depicts the technical framework for the experimental data capture and workflow management.

There are five functionalities offered to conservators and scientists after a successful login to the Experimental Data Capture system. The following list provides information on each of these functionalities:

- **Experimental Data Capture**: this interface allows users to create records, add metadata (samples metadata and observations) and characterisation images. Metadata are stored in the OPPRA-based knowledge-base – for inferencing and indexing. The characterisation images (e.g., Spectra, X-ray, cross-section images) are stored in the Jackrabbit repository – for a secure but accessible URL);

- **Workflow Management**: this is the overall user interface (explained in Chapter 6) that enables users to easily interact with all the functions available in the experimental data capture framework (e.g., create, search, publish, share and re-use) at any stage of the documentation process;
- **Search Experiments**: this is the search interface that allows users to search and view published experiments – via three available options: keyword search, ontology-based search, and SPARQL-based search;
- **Visualise Experiments**: this interface allows users to view the search results (e.g., records and workflows) using graph-based visualisation tools (e.g., RDF graphs);
- **Related Records/Publications**: this is the experimental data linking to past publications/experiments – which allows users to discover and compare similar/related experiments published via the 20th Century in Paint platform.

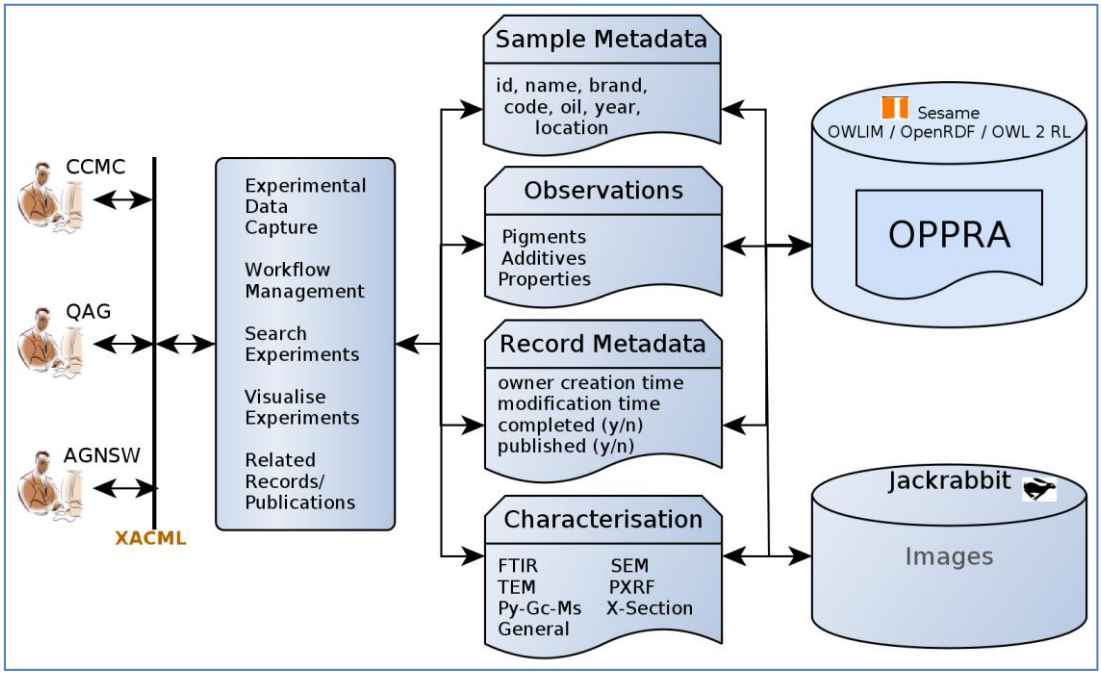



**Figure 5.4: Technical framework for the experimental data capture and workflow management**

### 5.2.5. Structured Data Extraction from Past Publications

The structured data extraction from past publications (*text2triples*) enables collaborators within the 20th Century in Paint project to capture and record the hidden knowledge existing in text publications in a form that complies with the OPPRA ontology. Figure 5.5 depicts the technical framework for the structured data extraction from past publications.

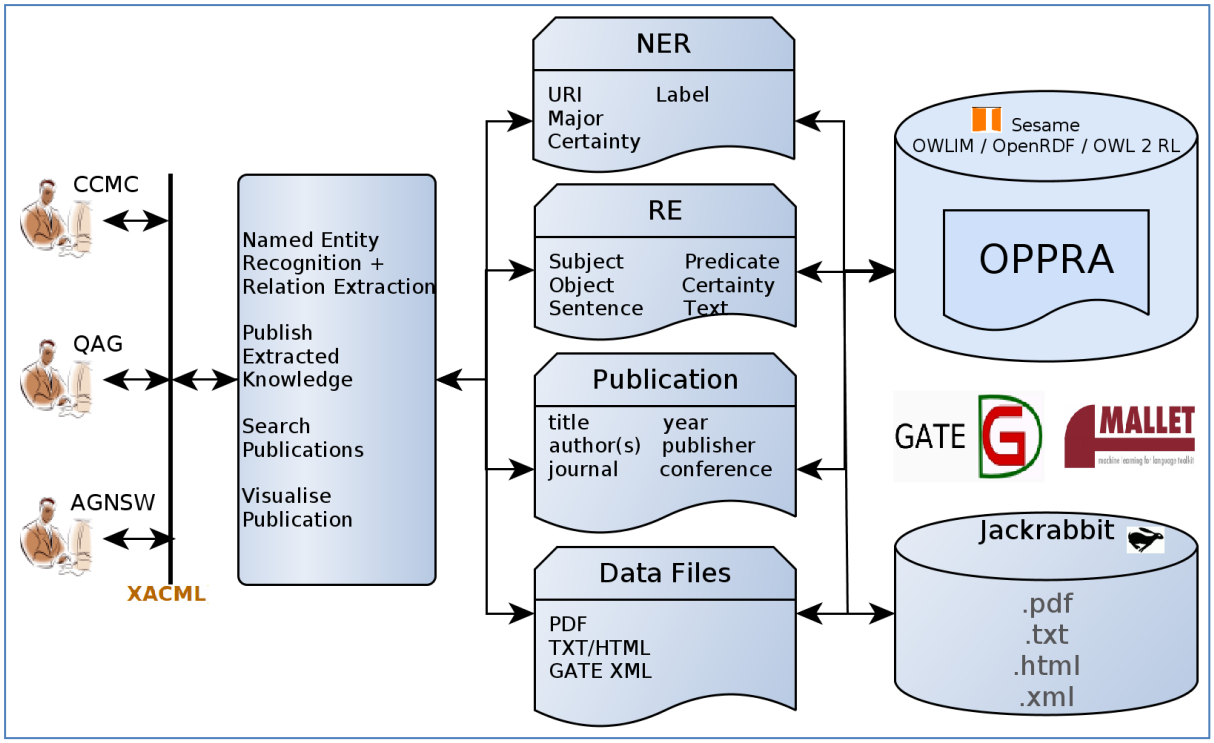



**Figure 5.5: Technical framework for the structured data extraction from past publications (*text2triples*)**

There are four functionalities offered to conservators and scientists after a successful login to *text2triples*. The following list provides information on each of these functionalities:

- **NER and RE**: this is the main interface that allows users to view automatically extracted entities and relationships (from the uploaded documents based on OPPRA), and enables them to modify and delete incorrectly extracted data;
- **Publish Extracted Knowledge**: this interface allows users to export the extracted triples into different formats (e.g., RDF/XML, Turtle, N3, TriX and TriG), to publish the triples – for indexing and reasoning, and to save the publications (into the Jackrabbit repository) for corpora building and future access/modification;
- **Search Publications**: this is the search interface that allows users to search and view publications – via three available options: keyword-based search, ontology-based search, and SPARQL-based search;
- **Visualise Publication**: this interface allows users to view the search results (e.g., publication – metadata, sentences and triples) using graph-based visualisation tools (e.g., RDF graphs).

### 5.2.6. Data Integration, Querying, Retrieval and Visualisation

The data integration, querying, retrieval and visualisation component was implemented in the 20th Century in Paint framework as '*Data Aggregation and Linking Interface – DALI*'. DALI enables collaborators to seamlessly search and visualise local and external knowledge (from experiments, publications and external databases). Figure 5.6 depicts the technical framework for DALI.

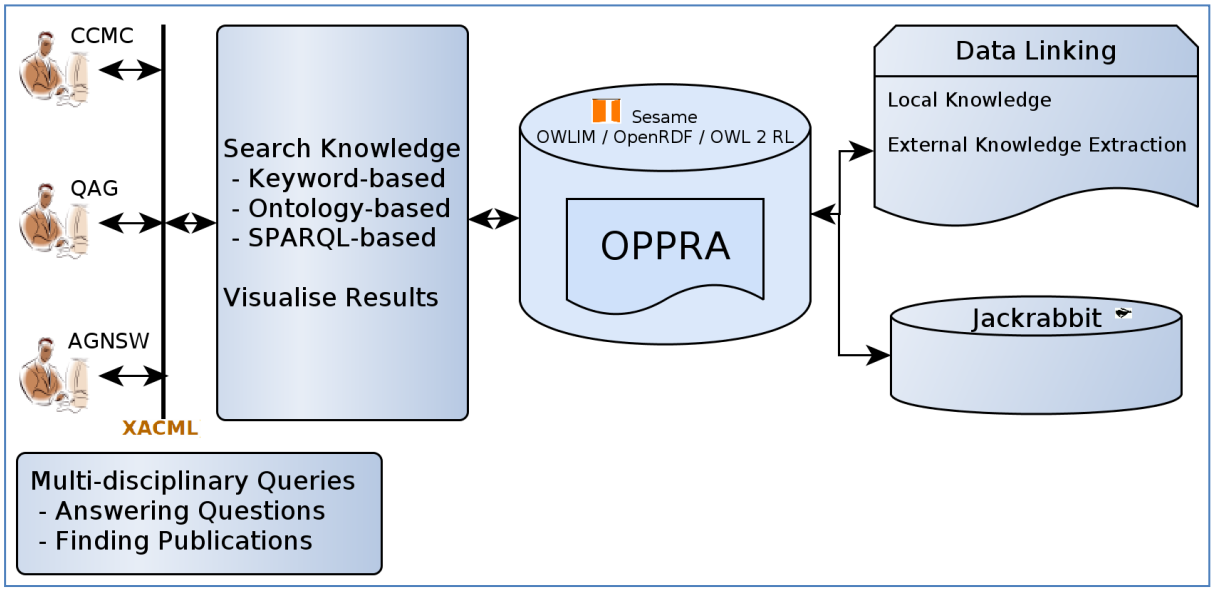



**Figure 5.6: Technical framework for the Data Aggregation and Linking Interface (DALI)**

There are two functionalities offered to conservators and scientists after a successful login to DALI. The following list provides information on each of these functionalities:

- **Search Knowledge**: this is the main interface that allows users to search the knowledge existing in the OPPRA RDF triple store. The interface offers keyword, ontology and SPARQL-based searches across the local data and external data, as well as the new knowledge obtained from the OWL 2 RL inferencing mechanism. Supported search queries can be divided into two forms: answering questions (e.g., what is the best solvent that will remove surface varnish from the painting *Epiphany*?), and finding references (e.g., list records/publications that provide information on the SEM characterisation of samples taken from Sidney Nolan paintings);
- **Visualise Publication**: this interface allows users to view the search results (e.g., publication – metadata, sentences and triples) using graph-based visualisation tools (e.g., RDF graphs).

## 5.3. Summary

This chapter documented the technical framework for the community-driven knowledge curation platform for the art/paint conservation domain. More specifically, it described the following components (and how they fit into the 20th Century in Paint framework):

- The Web portal, menu and Wiki (public and members' areas);
- Security aspects (authentication and access control);
- The OPPRA ontology, inferencing rules, the underlying triple store and the ontology browser;
- The user interface for the services described in Chapters 6-8 (experimental data capture, structured data extraction from past publications, and data aggregation and linking interface).

# Chapter 6

# Storing, Searching, Retrieving and Visualising Experimental Data

## 6.1. Introduction

Data management has become a critical challenge faced by the discipline of art/paint preservation in which the provision of conservation data management is pivotal to the achievements and impact of research projects (Green and Mustalish, 2009). Massive and rapidly expanding amounts of conservation data, combined with data models that evolve over time, contribute to making data management an increasingly challenging task that warrants a new approach.

Data management is the practice of managing (digital) data and resources, encompassing a wide range of activities including acquisition, storage, retrieval, discovery, access control, publication, integration, curation and archival (Gray et al., 2005). Historical and scientific data management informs and enables research within the art/paint preservation domain, of which it has become an indispensable component.

A workshop held for APTCCARN members at the Art Gallery of NSW in February 2010 determined that the data generated during the 20$^{th}$ Century in Paint project timeline potentially holds significant value both to its owner/creator and to other art

conservation researchers. One of the key aims of this workshop was to identify the data management, sharing and analysis requirements for the 20th Century in Paint project teams and their corresponding case studies (as described in Chapter 3). The key requirements identified at this workshop included:

- The ability to efficiently (and easily) acquire, store and manage large volumes of data;
- The ability to collaboratively add data and observations at any stage of the documentation process;
- The ability to maintain sufficient contextual information (conceptual domain models such as how research activities are organised and carried out; and metadata such as provenance information) – for more effective organisation, understanding, discovery, access, share and re-use of raw data;
- The ability to ensure data security through the use of authentication and authorisation solutions – including access control and archival;
- The ability to link publications to documented experiments (i.e., experiments and case studies that have been documented, and included in future publications) – through a persistent and unique naming scheme such as Digital Object Identifier – DOI and Named Graphs;
- The ability to search (and browse) for experiments and publications through mechanisms such as full-text searching, faceted browsing, complex query answering, and graph similarities;
- The ability to integrate, re-use and visualise both experimental data and provenance information.

For example, as seen in the case studies provided in Chapters 1 and 3, conservators and materials scientists perform various tasks for characterising paint materials (e.g., pigments, media and additives) before and after problems arise – to determine the optimum conservation treatments and environmental variables for storage, display and transport. Such activities generate a massive amount of historical and scientific data that needs to be efficiently managed. Examples of performed tasks and associated data include:

- Preparing mixtures – crm:E57.Material, oreChem:Quantity, crm:52.Timespan, crm:E39.Actor;

- Acquiring samples – from oppra:Artifact, from crm:E57.Material, crm:E53.Place, crm:E39.Actor;
- Experiments – crm:E55.Type, crm:E57.Material, crm:52.Timespan, crm:E39.Actor, crm:E53.Place;
- Condition assessments – rdfs:comment, oppra:ConditionReport, crm:52.Timespan, crm:E39.Actor, crm:E53.Place;
- Art history analysis – oppra:Artist, oppra:Genre, oppra:OilPaint, crm:E54.Dimension, crm:E53.Place;
- Characterisations – oppra:CharacterizationTechnique, oreChem:Instrument, oppra:calibration, oppra:outputs;
- Used materials – oppra:PhysicalAttribute, oppra:ChemicalAttribute, oppra:ChemicalReaction, oppra:ConditionState;
- Actors performing these tasks – oppra:Artist, oppra:Conservator, oppra:Scientist, oppra:Manufacturer, oppra:ArtGallery, oppra:Museum.

Database systems have traditionally been used successfully to manage research data (Shah et al., 2007) in which database schemas are used as domain models to capture the attributes and relationships of domain concepts. One implication of this approach is that the domain models need to stay relatively stable as database extension and migration is often an error-prone and laborious task. Consequently, this approach is not suitable for domains where data and model evolution is the norm rather than the exception.

Semantic Web ontology languages such as RDF Schema and OWL possess expressive, rigorously-defined semantics and non-ambiguous syntaxes. Moreover, they have been designed to be open and extensible and to support knowledge and data exchange on the Web (Auer et al., 2007, Berners-Lee, 2009). These intrinsic characteristics make them an ideal conceptual platform on which a flexible data management system for the art/paint preservation domain can be built. ***The hypothesis in this thesis is that Semantic Web technologies can facilitate an efficient experimental data storage approach for 20$^{th}$ century art/paint conservation – by enabling experiments (and provenance data) to be documented, linked, searched, discovered, shared, re-used and visualised.***

This chapter discusses the Experimental Data Capture platform. The framework provides an ontology-based experimental process system that enables conservators and materials scientists to collaboratively capture, store, share, link and compare historical and experimental data associated with 20th century art/paint preservation. Based on the OPPRA ontology (described in Chapter 4), the captured information will be stored in an RDF triple store where it can be searched and re-used by art conservators and scientists.

## 6.2. Related Work

A variety of projects have applied semantic technologies to capture (store and index) experimental data for e-Science applications including: Cultural Heritage (Baruzzo et al., 2008, Challapalli et al., 2006, Haslhofer et al., 2010, Hohmann and Schiemann, 2013, Sanderson and Van de Sompel, 2010, Schmidt et al., 2011), chemistry (Krafft et al., 2010, Pirró et al., 2010, Reid and Edwards, 2009), and natural sciences (Abidi et al., 2012, Smith et al., 2011). These projects cover a range of disciplines, but share a common desire for rich semantic querying capabilities for their data collections. Examples of the application of Semantic Web technologies to the capture (and sharing) of experimental data resulting from e-Science research activities include: E-Dvara; WissKi; Europeana; ourSpaces Virtual Research Environment; Vivo ontology-based approaches and DOKMS; Scratchpads; and the Platform for Ocean Knowledge Management (POKM); each of which is briefly described below.

*E-Dvara*

The E-Dvara project (Baruzzo et al., 2008, Challapalli et al., 2006) focused on the development of a new platform for the storage of digital contents (by integrating an RDF semantic layer) for Indian cultural heritage. It was designed to: a) reduce the effort required by the archivist to define the data structure used to represent data into the archives; b) provide (to archivists with no expertise in data management) a set of wizards devoted to data schemata creation; c) allow content providers to easily share their archives on the Web; and d) allow archivists to provide a specific visualisation template and a set of search forms.

*WissKi*

WissKi (Hohmann and Schiemann, 2013) within the Scientific Communication Infrastructure project is a semantic Wiki application (implementing the CIDOC-CRM) to work on use-cases in art history (e.g., Goldsmith's Art in Nuremberg, 16th-19th century) as well as natural history (e.g., the research diaries of a famous 19th century entomologist, Wilhelm Aerts). The data management infrastructure in the project uses Drupal for content management (including files) with the following extensions: OWL/RDF system, discussion system, automatic text annotator, authority files management, import/export API, and ARC2 triple store.

*Europeana*

The Europeana project (Haslhofer et al., 2010, Sanderson and Van de Sompel, 2010, Schmidt et al., 2011) aims to provide a search platform that integrates a collection of European digital libraries with digitised paintings, books, films and archives. The content creation used in Europeana enables archivists to annotate digital documents (e.g., videos, and maps) from the project, and link them to external resources, and enables the robustness of the annotations to be focused on over time by combining the temporal features built into the emerging Open Annotation model.

*OurSpaces Virtual Research Environment*

The aim of the OurSpaces Virtual Research Environment (Reid and Edwards, 2009) is to allow structured semantic metadata to interoperate with community-driven metadata (e.g., social networking, digital resource management such as upload, search and annotation, creation of project “spaces” to manage membership and activities according to project stages, privacy control; the publishing of blogs and wikis; and execution/monitoring of workflows). It was constructed using aspects of the Social and Semantic Web such as: OWL-Lite ontologies for *Agent*, *ScientificResource* and *ResearchTask*; Sesame RDF triple store for metadata storage and indexing; as well as the myExperiment Virtual Research Environment for workflow publishing.

*Vivo ontology-based approaches and DOKMS*

More recently, ontology-based approaches have been taken in Vivo (Krafft et al., 2010) to model, organise and integrate research activities and researcher profiles in

an institutional setting. DOKMS (Pirró et al., 2010) is a distributed, ontology-based knowledge management framework that serves similar purposes as Vivo. DOKMS operates in a P2P environment with desktop clients instead of browser-based as in Vivo.

*Scratchpads*

The Scratchpads project, developed by researchers and IT specialists of the Natural History Museum in London (Smith et al., 2011), is built on Drupal and allows for the creation, management, sharing, and publishing of taxonomic and other biodiversity information. It provides an integrated workbench and collaborative, open access space for the research community, using a number of modules and services that allow users to work on taxonomic classifications and import or link to specimen records, images, maps and literature.

*Platform for Ocean Knowledge Management – POKM*

POKM (Abidi et al., 2012) enables researchers to design and execute complex experiments by composing specialised experimental workflows that are suited for their scientific tasks. The semantic framework allows the following services: a) the selection and sharing of multi-modal data; b) the visualisation of multiple data layers at a geographic location; c) the interconnection of different research communities so that they can seamlessly interact and share data, scientific models, experiment results, knowledge resources and expertise; and d) the cataloguing of experiment-specific data and knowledge so that it can be used for future experiments.

In addition, several domain-specific online workflow repositories (with semantic extensions) have evolved in recent years. For example, Kepler (Altintas et al., 2006, Ludäscher et al., 2006, Mcphillips et al., 2006) provides a user-friendly interface that supports the design and re-use of Grid workflows. It is designed with advanced features for composing and accessing local and remote scientific applications. Taverna (Hull et al., 2006, Oinn et al., 2004) is a high-level middleware for supporting bioinformatics workflows. It provides data models, enactor task extensions and graphical user interfaces with state transition and multithreading mechanisms to speed up the data acquisition process. myExperiment (De Roure et al., 2007, De Roure et al., 2009, Goble et al., 2006, Goble and De Roure, 2007) is a

social Website sharing scientific workflows and research objects (e.g., ratings, metadata such as tag clouds, papers and other workflows including Kepler and Taverna projects). It has a REST interface to access its publicly available dataset.

The projects reviewed above aim to improve the discoverability of experimental data by capturing and integrating heterogeneous digital collections via rich semantic metadata and common models. They don't, however, focus on the specific requirements associated with the conservation of cultural artefacts, as outlined in a workshop on conservation held by the US National Science Foundation in 2009 (Leona and Duyne, 2009). For example, they lack the following aspects with respect to the art/paint preservation domain:

- Fine-grained relationships between: a) events (intentional activities or unintentional processes) and their effects on cultural artefacts (condition changes), such as environmental factors (e.g., flood) and breakage and deterioration (e.g., discolouration) triggering another form of deterioration (e.g., yellowing), chemical reactions and oxidation, transportation and damage; b) physical provenance (e.g., transportation, exhibition, acquisition, assessment, cleaning) and digital provenance (e.g., sample, characterisation, identified materials, experiment); c) manufacturers/suppliers and materials; d) painting and artistic techniques; and e) characterisation activities and characterisation techniques;
- Semantic inferencing between relationships and terminologies such as: a) condition state describing condition change (e.g., "Blistered *describes* Blistering"; b) artist, painting and artistic techniques; c) characterisation, characterisation technique and instruments; d) manufacturer, manufacturing process and materials; and e) supplier, supplying process and materials;
- Recommendation systems that link published experiments, to other similar/related experiments in the literature.

Other projects on experimental data capture also exist within a range of semantic solutions, a key selection of which are discussed in Chapters 7-8.

## 6.3. Ontology-based Experimental Data Capture

Although this system is designed to be used by any conservator or materials scientist, the Experimental Data Capture (e.g., functionalities of data storage, linking

and retrieval) has been evaluated in the context of the 20th Century in Paint project. A series of integrated studies are being performed by the project's researchers focusing on art history and conservation, materials development and deterioration, and scientific tools and techniques (Chapter 3). The generated and shared information on paint samples within these studies includes:

- Paint information – name, brand, medium, year, code and sample location;
- Observations such as form (liquid/solid), colour and texture;
- Microscopic images such as FTIR, XRF, Py-Gc-Ms, SEM, and TEM;
- Identified materials such as pigments, additives and chemicals;
- Files associated with this analysis (e.g., XML, JPEG and SPA files).

The ontology-based Experimental Data Capture in this project manages the experimental workflow of the various tasks conducted by conservators and scientists – to allow them to upload their data and findings to the knowledge-base, share and re-use this data, and make the data accessible publicly. For example, Figure 6.1 illustrates an ontological representation of an experimental workflow that aims to investigate the impact of different environmental parameters on paint samples. It involves preparing samples of zinc oxide, fatty acids, additives and polymer (matrix), exposing them to different environmental conditions such as controlled temperatures and relative humidity over different time periods, and then analysing them (before and after exposure) using different characterisation techniques (e.g., Raman spectroscopy).

This graph (experimental data including classes, objects and relations) can then be expressed using a data model (OPPRA) so it can be searched and re-used. The following section provides information on the ontological data storage and linking of experiments.

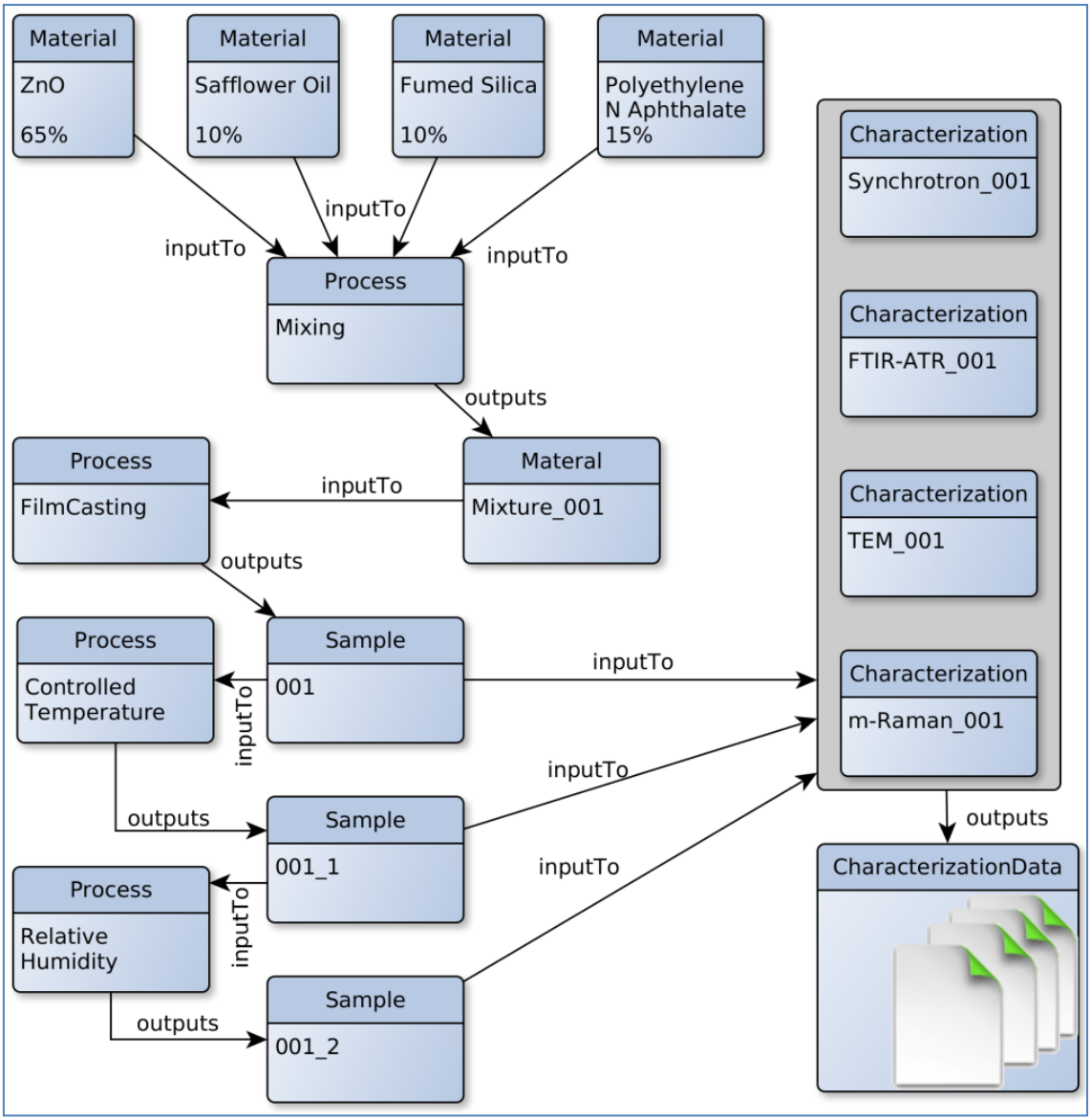



**Figure 6.1: Experimental representation that involves preparing samples of zinc oxide, fatty acids, additives and polymer (matrix), exposing and characterising them using different environmental and characterisation conditions**

### 6.3.1. The Ontology of Paintings and PReservation of Art – OPPRA

In this context, the OPPRA ontology (Chapter 4) is used for the ontology-centric modelling and processing. The OWL representation of OPPRA manages to effectively support a dynamic conceptual framework – in which: OWL classes represent the art/paint preservation domain concepts; OWL properties define concept attributes and their relationships; OWL restrictions specify constraints on concepts; and finally, OWL individuals define concrete art/paint preservation objects where attributes and relationships are defined using OWL assertions. Such a conceptual architecture alleviates the problem of imposing hard relational constraints in a database which is difficult to extend/change.

The key entities that are defined in the OPPRA ontology for the Experimental Data Capture system are: *oppra:Sample*; *oppra:Material* (where the sample is taken from such as Paint, or the material forming part of the sample such as binder, pigment, and additive); *oppra:Manufacturer* (the agent that manufactured the paint, if the sample was taken from a particular paint); *cidoc_crm:Timespan* (the date of the compound creation); *cidoc_crm:Activity* (the process that each artefact undergoes such as controlled temperature, assessment, and characterisation); *oppra:CharacterizationTechnique* (e.g., SEM); *cidoc_crm:Document* (output file as a result from a specific activity such as spectra); and *oppra:Record* (URI used to define the record's URL, database, and owner, and to contain the sample/experiment statements). Chapter 4 (Sections 4.3.1 and 4.3.2) illustrates how these key entities are connected through *owl:ObjectProperty* (for Entity/Entity relations), and *owl:DatatypeProperty* (for Entity/Value assignments).

Each record (*oppra:Record*) in each database represents a container (Named Graph) holding all the information about a specific sample. The following TriG exports contain N3 statements for the example given in Figure 6.1. The example represents Named Graphs (specific records) about the activity "Mixing Activity 001", and the three samples "Sample 001", "Sample 001_1", and "Sample 001_2". The information included in each record represents the core triples (about the given artefact) that are created by the Experimental Data Capture system. The *oppra:Record* enables experiments to be linked to related publications (as will be seen in Section 6.4.4). It also expedites access and re-use of data: via its specified URL (e.g.,oppra:url); via inferencing (e.g., *?database oai_ore:aggregates ?record . graph ?record{?s ?p ?o}*); and via SPARQL queries (e.g., *select distinct ?s ?p ?o from oppra:Mecklenburg_Record001 where {?s ?p ?o }*).

```
oppra:Mixing_Record_001{
  oppra:FumedSilica_10
    a owl:NamedIndividual, cidoc_crm:Material;
    oppra:qty "10%";
    oppra:inputTo oppra:Mixing_001 .
  oppra:Mixing_001
    a owl:NamedIndividual, cidoc_crm:Activity;
    oppra:outputs oppra:Mixture_001 .
  oppra:Mixture_001
    a owl:NamedIndividual, cidoc_crm:Material .
  oppra:PolyethyleneNAQphthalate_15
    a owl:NamedIndividual, cidoc_crm:Material;
    oppra:qty "15%";
    oppra:inputTo oppra:Mixing_001 .
  oppra:SafflowerOil_10
    a owl:NamedIndividual, cidoc_crm:Material;
    oppra:qty "10%";
    oppra:inputTo oppra:Mixing_001 .
  oppra:ZnO_65
    a owl:NamedIndividual, cidoc_crm:Material;
    oppra:qty "65%";
    oppra:inputTo oppra:Mixing_001 .
}

oppra:Sample_Record_001_1>{
  oppra:ControlledTemperature_001
    a owl:NamedIndividual, cidoc_crm:Activity;
    oppra:outputs oppra:Sample001_1 .
  oppra:FTIR_ATR_001_1
    a owl:NamedIndividual, oppra:Characterization;
    oppra:outputs oppra:FTIR_ATR_File_001_1 .
  oppra:MRaman_001_1
    a owl:NamedIndividual, oppra:Characterization;
    oppra:outputs oppra:MRaman_File_001_1 .
  oppra:RelativeHumidity_001_1
    a owl:NamedIndividual, cidoc_crm:Activity;
    oppra:outputs oppra:Sample001_2 .
  oppra:Sample001_1
    a owl:NamedIndividual, oppra:Sample;
    oppra:inputTo
      oppra:FTIR_ATR_001_1,
      oppra:MRaman_001_1,
      oppra:RelativeHumidity_001_1,
      oppra:Synchrotron_001_1,
      oppra:TEM_001_1 .
  oppra:Sample001_2
    a owl:NamedIndividual, oppra:Sample .
  oppra:Synchrotron_001_1
    a owl:NamedIndividual, oppra:Characterization;
    oppra:outputs oppra:Synchrotron_File_001_1 .
  oppra:TEM_001_1
    a owl:NamedIndividual, oppra:Characterization;
    oppra:outputs oppra:TEM_File_001_1 .
}

oppra:Sample_Record_001_2>{
  oppra:FTIR_ATR_001_2
    a owl:NamedIndividual, oppra:Characterization;
    oppra:outputs oppra:FTIR_ATR_File_001_2 .
  oppra:MRaman_001_2
    a owl:NamedIndividual, oppra:Characterization;
    oppra:outputs oppra:MRaman_File_001_2 .
  oppra:Sample001_2
    a owl:NamedIndividual, oppra:Sample;
    oppra:inputTo
      oppra:FTIR_ATR_001_2,
      oppra:MRaman_001_2,
      oppra:Synchrotron_001_2,
      oppra:TEM_001_2 .
  oppra:Synchrotron_001_2
    a owl:NamedIndividual, oppra:Characterization;
    oppra:outputs oppra:Synchrotron_File_001_2 .
  oppra:TEM_001_2
    a owl:NamedIndividual, oppra:Characterization;
    oppra:outputs oppra:TEM_File_001_2 .
}

oppra:Sample_Record_001>{
  oppra:ControlledTemperature_001
    a owl:NamedIndividual, cidoc_crm:Activity .
  oppra:FTIR_ATR_001
    a owl:NamedIndividual, oppra:Characterization;
    oppra:outputs oppra:FTIR_ATR_File_001 .
  oppra:FilmCasting_001
    a owl:NamedIndividual, cidoc_crm:Activity;
    oppra:outputs oppra:Sample001 .
  oppra:MRaman_001
    a owl:NamedIndividual, oppra:Characterization;
    oppra:outputs oppra:MRaman_File_001 .
  oppra:Mixture_001
    a owl:NamedIndividual, cidoc_crm:Material ;
    oppra:inputTo oppra:FilmCasting_001 .
  oppra:Sample001
    a owl:NamedIndividual, oppra:Sample;
    oppra:inputTo
      oppra:ControlledTemperature_001,
      oppra:FTIR_ATR_001,
      oppra:MRaman_001,
      oppra:Synchrotron_001,
      oppra:TEM_001 .
  oppra:Synchrotron_001
    a owl:NamedIndividual, oppra:Characterization;
    oppra:outputs oppra:Synchrotron_File_001 .
  oppra:TEM_001
    a owl:NamedIndividual, oppra:Characterization;
    oppra:outputs oppra:TEM-File_001 .
}
```

## 6.4. System Implementation and User Interface

The Experimental Data Capture system is designed to support the following objectives:

- To operate as a Web-based experimental workflow system that enables collaborators (e.g., within the 20th Century in Paint project) to quickly and easily describe and publish their experiments and data;
- To capture new information in a form that complies with the OPPRA ontology;
- To enable experimental data and observations to be added (securely) at any stage of the documentation process;
- To enable experiments to be linked to future publications;
- To enable experiments to be searched and browsed based on different search mechanisms (e.g., keyword-based search, ontology-based search, and graph similarities);
- To enable experiments to be integrated, re-used and visualised (e.g., RDF exports, and graph visualisations).

The following sections provide details on the system architecture, as well as the user interface for capturing and publishing experimental data, searching and visualising experimental data, linking experiments to publications, and searching and retrieving similar experiments.

### 6.4.1. System Architecture

Figure 6.2 illustrates the overall architecture and major components of the experimental data management system.

The Web-based framework is implemented within the 20th Century in Paint project's website, and divided into different workspaces (e.g., Sidney Nolan Paint Archive, and Mecklenburg Samples). It provides a single user interface to the locally deployed storage and services. AJAX, JSP, JavaScript and CSS are the underlying technologies – chosen to ensure dynamic Web interfaces and highly responsive interactivity. The Web portal has been developed using a combination of: Apache Tomcat, MySQL, Apache Jackrabbit and the Sesame RDF triple store.

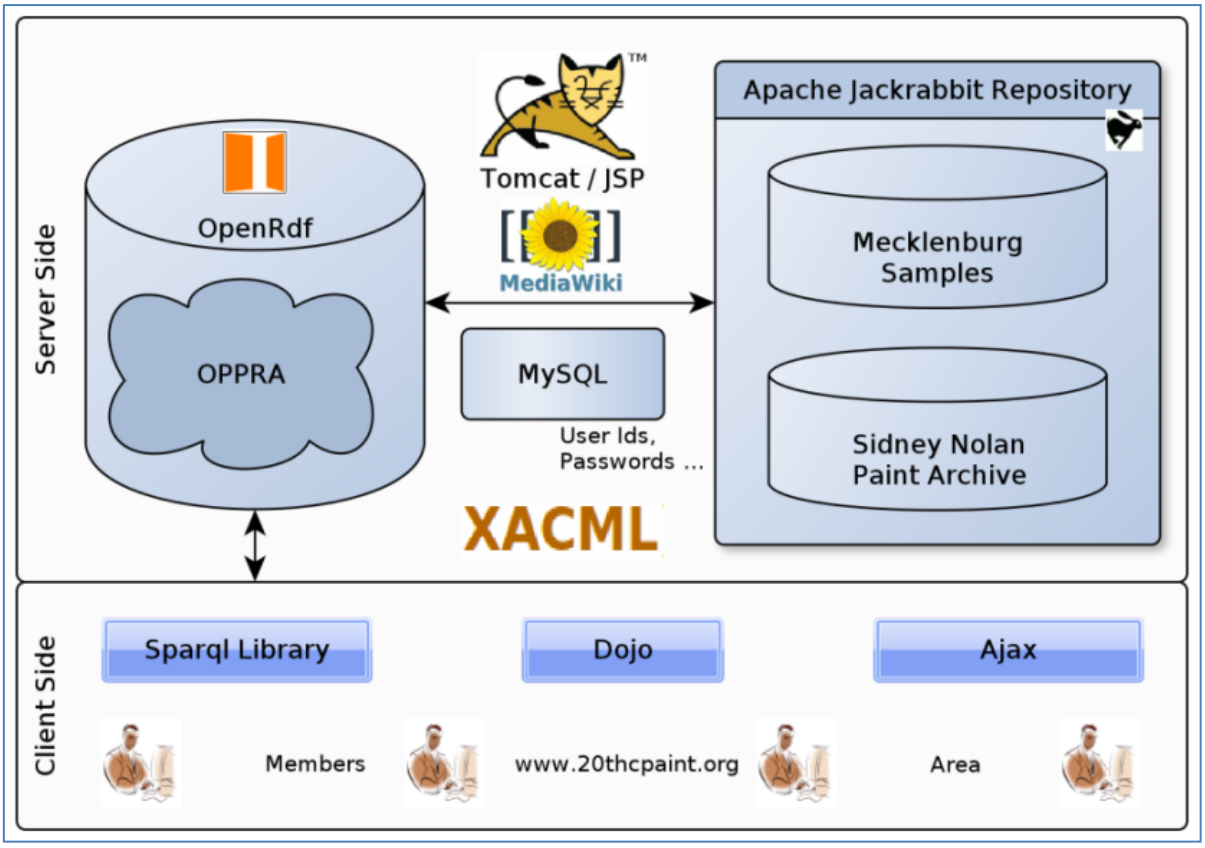



**Figure 6.2: Overall architecture and major components of the Experimental Data Capture platform**

Apache Jackrabbit is used as the content repository for storing images and spectra (Apache, 2004). Apache Jackrabbit offers support for multiple pluggable storage back-ends, fast data modifications, the ability to associate metadata with different file formats, and the XPath-enabled mechanism to search for files and content. Its security features are extendable to work with fine-grained access controls such as XACML.

The OPPRA-compliant RDF instances (as described above) are stored in an OpenRDF repository – the Sesame triple store (Aduna, 1997). The metadata (in the RDF triple store) contain links to the image and spectra files stored in the Jackrabbit repositories. The local RDF triple store and associated Jackrabbit repositories are only accessible to authenticated project members.

### 6.4.2. Capturing and Publishing Experimental Data

The user interface for capturing and publishing experiments allows users to create a sample record with metadata including: id (e.g., oppra:Sample007 *oppra:hasId* "007"), brand (e.g., oppra:Sample007 oppra:takenFrom *oppra:Grumbacher*), name (e.g., oppra:Grumbacher *oppra:hasName* Grumbacher), and binding medium (e.g., oppra:Grumbacher *oppra:containsMaterial* oppra:LinseedOil). Once a sample is created, its information can be edited by authorised users (except the id field). Figure 6.3 shows the HTML form for the sample record creation.

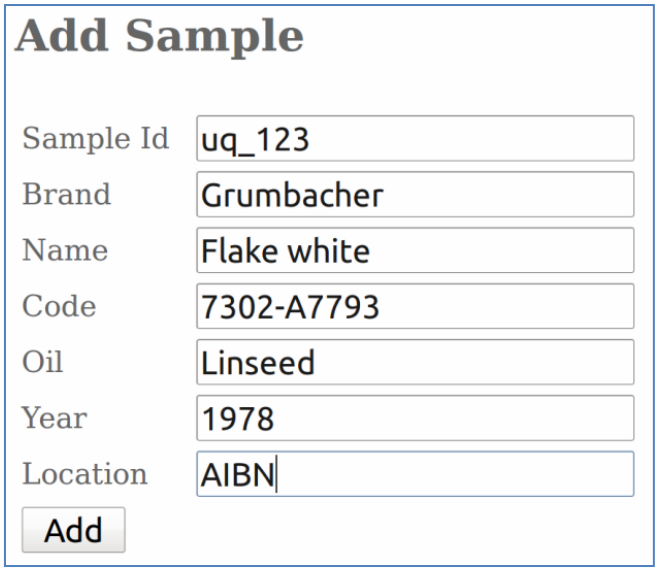



**Figure 6.3: Sample record creation**

Observation and characterisation data can then be uploaded and edited accordingly. Figure 6.4 shows a complete example of sample information entered by a Team 2 member into the Mecklenburg Samples repository.

Each record allows conservators and scientists to perform additional tasks to help them track changes and share information on that particular sample (or experiment). For example, the user may toggle the status of each record (e.g., completed, not completed, published and unpublished). They can also delete, undelete, restore and permanently delete any set of records at any time in the workflow process. Finally, they can see the edit types of the selected record (e.g., sample creation, file upload and observation addition), as well as the edit users and dates. Figures 6.5-6.7 show snapshots of various tasks previously performed on a set of Sidney Nolan Paint Archive records.

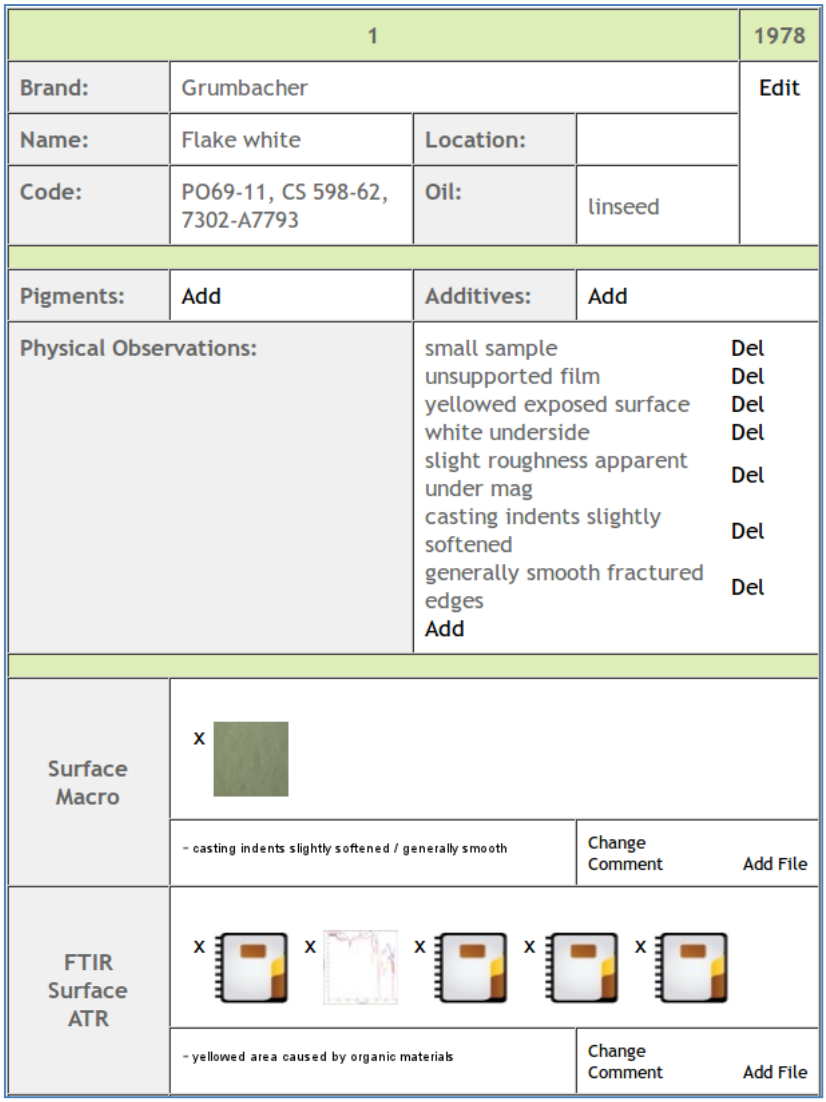



**Figure 6.4: Sample record showing basic information and characterisation data**

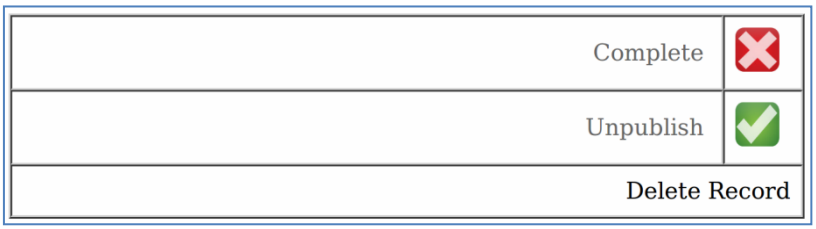
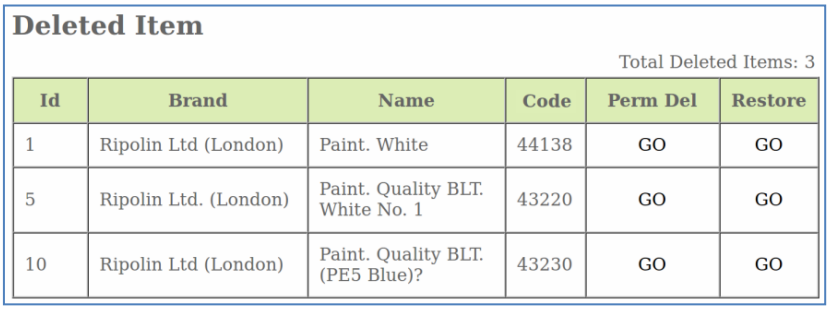



**Figure 6.5: Set a record as completed, published and/or deleted**

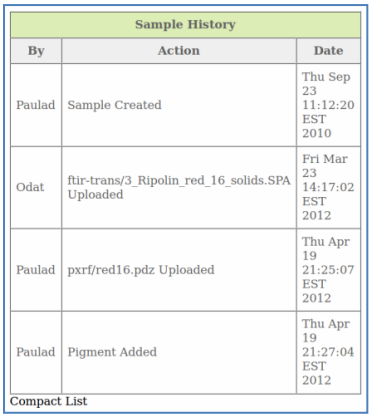


**Deleted Item**

Total Deleted Items: 3

| Id | Brand | Name | Code | Perm Del | Restore |
|---|---|---|---|---|---|
| 1 | Ripolin Ltd (London) | Paint. White | 44138 | GO | GO |
| 5 | Ripolin Ltd. (London) | Paint. Quality BLT. White No. 1 | 43220 | GO | GO |
| 10 | Ripolin Ltd (London) | Paint. Quality BLT. (PE5 Blue)? | 43230 | GO | GO |

**Figure 6.6: Restore or permanently delete a record that was set to be deleted**



Sample History

| By | Action | Date |
|---|---|---|
| Paulad | Sample Created | Thu Sep 23 11:12:20 EST 2010 |
| Odat | ftir-trans/3_Ripolin_red_16_solids.SPA Uploaded | Fri Mar 23 14:17:02 EST 2012 |
| Paulad | pxrf/red16.pdz Uploaded | Thu Apr 19 21:25:07 EST 2012 |
| Paulad | Pigment Added | Thu Apr 19 21:27:04 EST 2012 |

Compact List

**Figure 6.7: Snapshot showing a series of edits performed on a record**

### 6.4.3. Searching and Visualising Experimental Data

The RDF representation of published data makes it easy for human and computer agents to perform queries across the knowledge-base. To do this, the search interface takes advantage of the SPARQL-based REST API offered by OpenRDF.

One of the main challenges of SPARQL is that querying databases poses difficulties to non-technical users. A user-friendly search interface is implemented for the Experimental Data Capture platform that enables non-technical users to perform keyword and OPPRA-based queries across the stored data. Figure 6.8 shows the interface and results of the query: "*Give me all FTIR-ATR characterisations that were performed on materials containing zinc oxide.*"

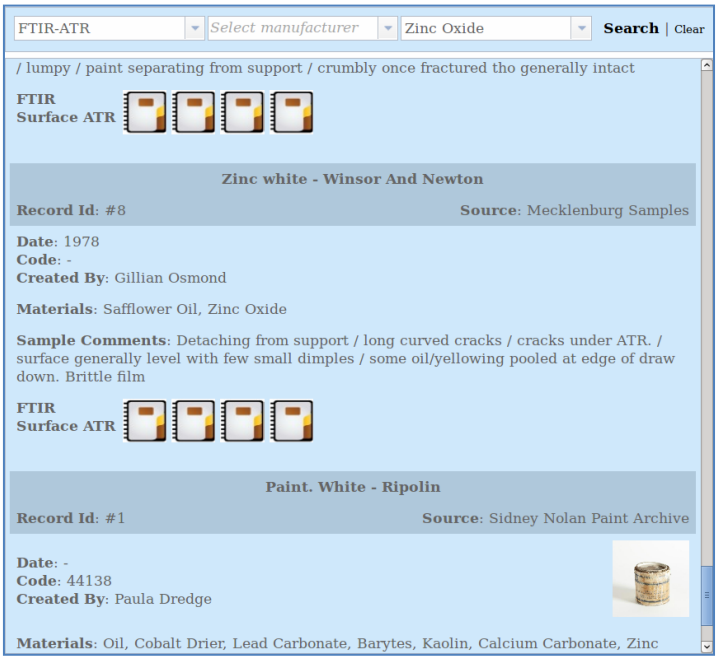



**Figure 6.8: Interface and result of a query "show all characterisations of materials containing zinc oxide"**

The resulting graphs (SPARQL constructs) can also be manipulated to allow for different visualisation tools. For example, Figure 6.9 shows the InfoVis tool (Belmonte, 2013) displaying information on an RDF graph of Sidney Nolan Archive record 'SampleRecord6' (Ripolin Paint, Black No. 1105). The InfoVis tool takes a JSON string that was created by manipulating the SPARQL RDF result.

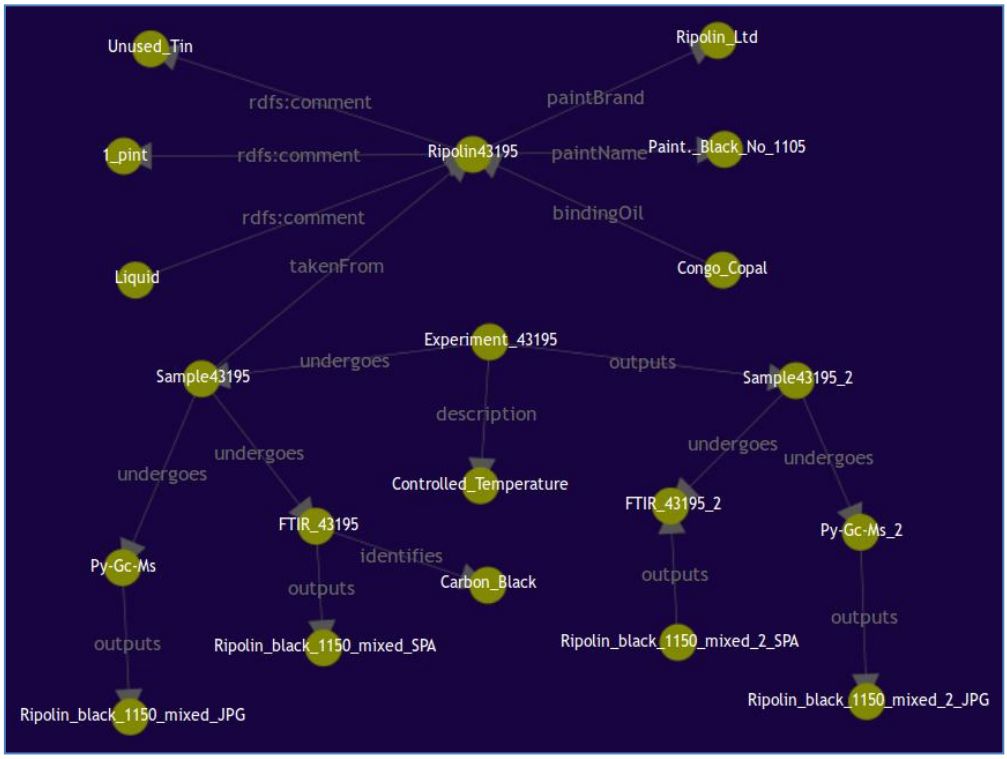



**Figure 6.9: InfoVis visualisation tool displaying SPARQL RDF result converted to JSON format**

### 6.4.4. Linking Experiments to Publications

The use of Named Graphs (OpenRDF contexts) enables an experiment to be:

- Linked to (and re-used by) publications using the transitive property *oai_ore:aggregates*. This property indicates that publications, and/or a set of sentences via inferencing, are referring to that experiment. Chapter 7 provides details on the structured data extraction (including experiments) from textual publications;
- Discovered by the SPARQL *construct*, *from*, and *graph* keywords. In addition to the SPARQL search discussed in Section 6.4.3, these keywords (*construct*, *from*, and *graph*) can query the knowledge-base based on a given full graph (e.g., other similar/related experiments). Graph-based search will be addressed in Sections 6.4.5 and 8.7.

### 6.4.5. Searching and Retrieving Similar Experiments

The ontology-based modelling and storage of experiments allow similar (and related) experiments to be searched and retrieved. The graph matching methodology adopted for discovering and ranking similar experiments is described as follows:

- The graph matching mechanism uses SPARQL queries – the SPARQL input is a *union* operation of all triples (subject, predicate, object) existing in the original graph;
- Inferencing rules are applied on both the original graph and the overall knowledge-base;
- The graph similarity (GS) between the original graph ($g_{org}$) and recommended graph ($g_{pred}$) is calculated as follows:

$$GS(g_{org}, g_{pred}) = \sum_{i=1}^{n} \frac{CS_{\mathrm{i}}(\mathrm{subj}_{\mathrm{org}}, \mathrm{subj}_{\mathrm{pred}}) + CS_{\mathrm{i}}(\mathrm{obj}_{\mathrm{org}}, \mathrm{obj}_{\mathrm{pred}})}{2 * n}$$

where,

$n$is the number of relations in the recommended graph that are shared with the original graph;

$\mathrm{subj}_{\mathrm{org}}$, $\mathrm{subj}_{\mathrm{pred}}$, $\mathrm{obj}_{\mathrm{org}}$, $\mathrm{obj}_{\mathrm{pred}}$ are the edges (of the $i^{th}$ predicate): subject in the original graph, subject in the recommended graph, object in the original graph and object in the recommended graph, respectively;

$CS_{\mathrm{i}}$ is the concept similarity between the $i^{th}$ predicate's edges (subject, object) – calculated using the *edge*-based semantic similarity (Wu and Palmer, 1994) that takes into account the paths (using $i^{th}$ predicate) between the concepts in the OPPRA ontology as follows:

$$CS(C_1, C_2) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3}$$

$N_1$is the length of the path from $C_1$ to the root;

$N_2$is the length of the path from $C_2$ to the root;

$N_3$is the length of the path from the most informative common ancestor to the root.

For example, suppose that a conservator "Gillian" enters information on a characterisation "FTIR-ATR" procedure that takes the sample "*uq_17*". The

characterisation identifies the presence of substances (lead carbonate, linseed oil and lead monoxide) and outputs three spectra files. When Gillian is ready to publish these results, the system detects slightly similar results that exist in the OPPRA-based RDF triple store (published via the *Experimental Data Capture, text2triples* discussed in Chapter 7, and *Data Extraction from External Databases* discussed in Chapters 6-8) – the system then notifies Gillian and gives her an option to open and inspect these publications (or experiments).

Figure 6.10 presents the original published graph (left) with two similar graphs that show samples being characterised by FTIR-ATR with properties such as materials identifications and characterisation outputs. The first graph (top-right) is from the Sidney Nolan Paint Archive with a similarity of 85.3%, and the second graph (bottom-right) is from a publication "*2007 Attenuated total reflection micro FTIR characterisation of pigment-binder interaction in reconstructed paint films*" with a similarity of 50%.

## 6.5. Evaluation

This section provides information on the setting and results of the experiment conducted to evaluate the effectiveness of linking experimental data to past publications.

### 6.5.1. Experimental Setting

The effectiveness of the linking task was evaluated using the following two standard metrics borrowed from Information Retrieval:

- **Precision@k** defined as the ratio between the number of relevant tags taken from the top-k recommended tags for record *i* and the number of tags considered;
- **Recall@k** defined as the ratio between the number of relevant tags in the top-k for record *i* and the total number of relevant tags.

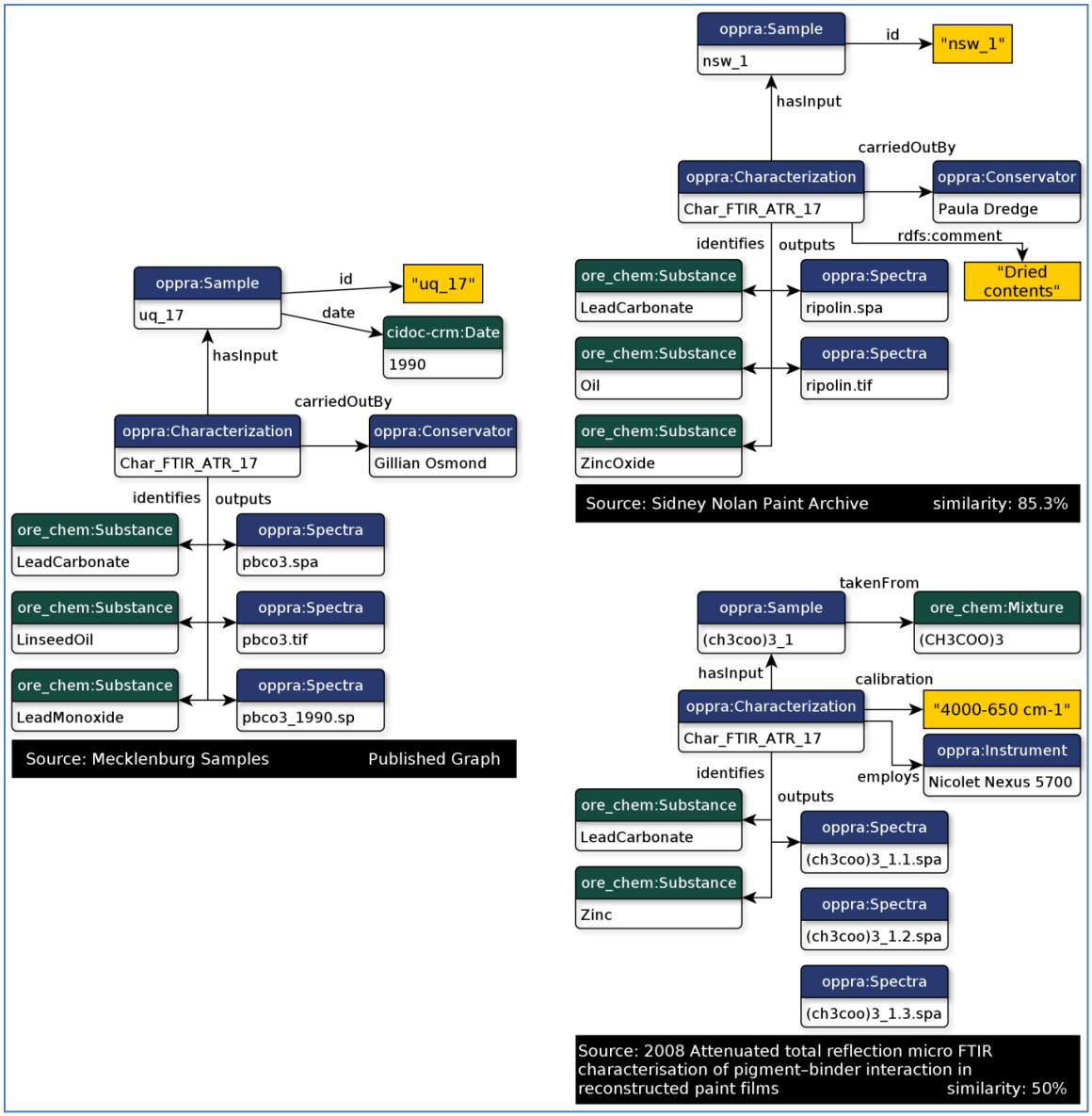



**Figure 6.10: Example of a published graph alongside similar graphs identified using the graph mechanism – similarity is 85.3% (top-right) and 50% (bottom-right)**

Real data available in the Jackrabbit repositories (20th Century in Paint local databases) were used for these experiments. The following list summarises the contents of these repositories:

- Datasets: Sidney Nolan Paint Archive and Mecklenburg Samples
- Number of Records: 160
- Number of sub-graphs: 374 – consisting of:
  - Paint/sample metadata (e.g., id, paint name, oil and year): 160
  - Identified materials (e.g., pigments, additives and chemicals): 72
  - Characterisation activities including file results and observations (e.g., FTIR, XRF, SEM and TEM): 140.

The document collection contained 100 publications about paint conservation from 15 different journals (e.g., Journal of the American Institute for Conservation, JSTOR Studies in Conservation, Analytical Chemistry, AICCM Bulletin, etc.). The 100 collected publications were selected and sourced with input from the 20th Century in Paint project teams. The 100 publications were selected on the basis of both quality (high impact factor and citations) and relevance to the case studies. Copyright issues were not a factor in this process because the documents were indexed/annotated for search/querying purposes only – links to the actual full-text document were provided if the user wished to proceed to read the document (but this would depend on his/her access rights via the publisher's Website). The documents were manually annotated using terms from the OPPRA ontology – the document annotation task is presented in more detail in Chapter 7. The training and testing documents were selected randomly in a 10-fold procedure, with a ratio of 75:25 (75% training, and 25% testing data).

The experimental data used for the evaluation is available (and can be queried) online – DALI (20thcpaint, 2012a).

### 6.5.2. Experimental Results

Four different techniques for the published experiments below were compared – based on term and edge weightings (Resnik, 1995, Salton and Buckley, 1988, Wu and Palmer, 1994) which are the key techniques used in most large-scale information retrieval systems:

- **Concept-based Graph Similarity** that calculates the similarity between the original and recommended graphs by adding the number of shared concepts in the recommended graph (divided by the maximum similarity obtained from all recommendations):

$$GS(g_{org}, g_{pred}) = \sum \frac{Shared\ Concepts}{\max(GS)}$$

- **Relation-based Graph Similarity** that calculates the similarity between the original and recommended graphs by adding the number of shared relations in the recommended graph (divided by the maximum similarity obtained from all recommendations):

$$GS(g_{org}, g_{pred}) = \sum \frac{Shared\ Relations}{\max(GS)}$$

- **Concept/Relation-based Graph Similarity** that calculates the similarity between the original and recommended graphs using the method introduced in Section 6.5. This method, however, does not consider the inferencing rules applied on the given graphs;
- **Concept/Relation-based Graph Similarity with Inferencing** that calculates the similarity between the original and recommended graphs using the method introduced in Section 6.5 – considering inferencing rules on the full dataset (including the original graph).

Figure 6.11 compares the four experimental techniques on a Precision@k results graph. Results for Recall@k are depicted on Figure 6.12. The observations of the performance of the techniques in the experiments are discussed below.

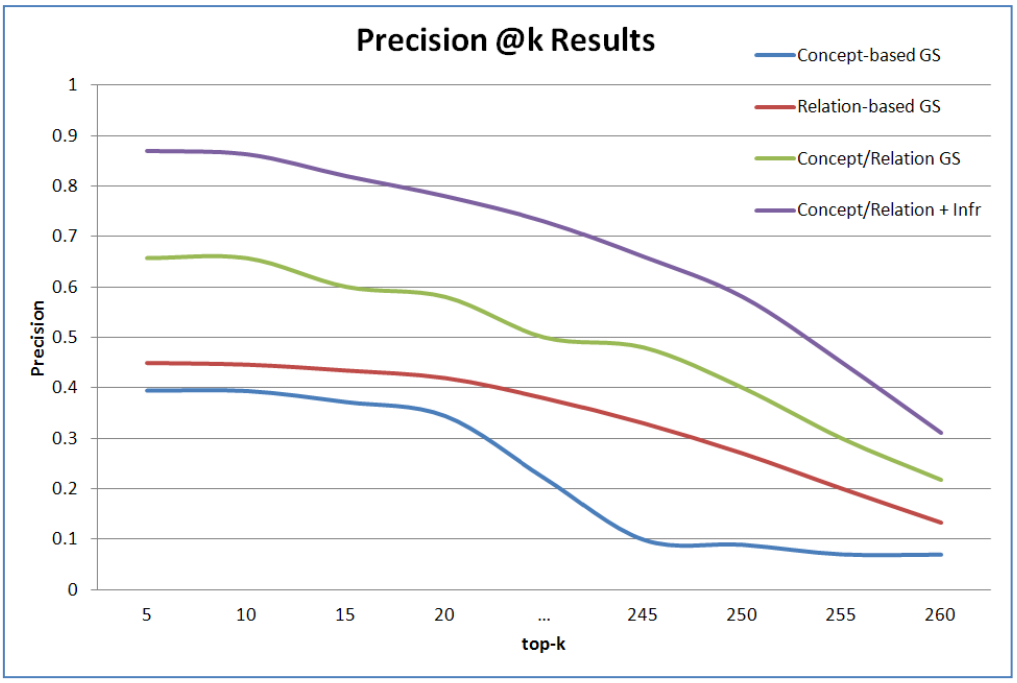



**Figure 6.11: Precision@k results for the four experimental techniques**

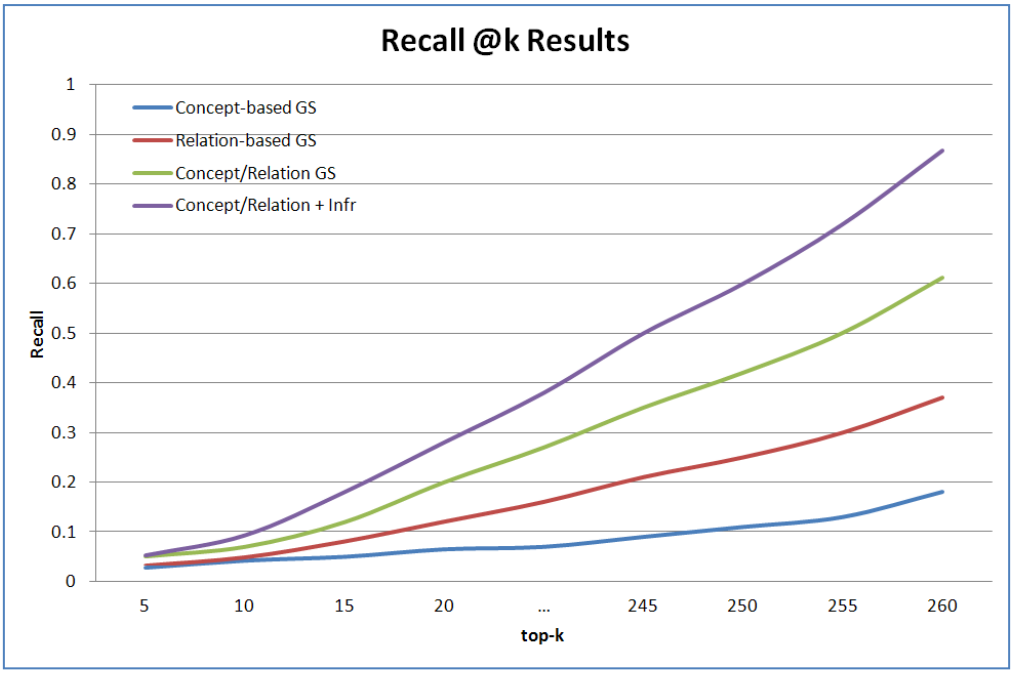



**Figure 6.12: Recall@k results for the four experimental techniques**

- The concept-based GS yielded the worst precision, starting from 39% when the top-k was 5, and it dropped dramatically after the top-k was 20 – reaching 8% for all results (top-k=260). On the other hand, the recall improved slowly as k increased – ranging from 3.5% to 18%;
- The relation-based GS improved the concept-based GS precision by 5%, but its trend stayed consistent in both precision (downward) and recall (upward) throughout the displayed results (from k=5 to k=260). The consistency of the observed trends was due to the generality of the existing relations' edges (the domain/range concepts surrounding each relation in the original graph);
- The concept/relation-based GS without inferencing improved the relation-based GS precision and recall by 20% in the first 4 experiments (from k=5 to k=20). However, it dropped inconsistently in the precision graph from that point due to the specificity of each required triple (e.g., subjects, predicates and objects existing in the original graph that needed to be found in the recommended graphs);
- The concept/relation-based GS with inferencing improved the relation/concept-based GS without inferencing precision by 20%. The downward trend in the precision experiment stayed consistent throughout the displayed results (from k=5 to k=260) due to the possible variants of the required triples (e.g., subjects, predicates and objects existing in the original graph, as well as the inferred

triples by OWL 2 RL that needed to be found). For the same reason, the recall results were dramatically improved as the top-k increased (1% for k=5, to 27% for k=260).

Figure 6.13 shows the precision/recall results for the four experiments. It shows that the precision/recall rate for the Concept/Relation method with inferencing, decreases more slowly and less, compared with the precision/recall decrease rates of the other methods (Concept, Relation, and Concept/Relation).The actual accuracy of the Concept/Relation-based GS with inferencing was impressive (87%).

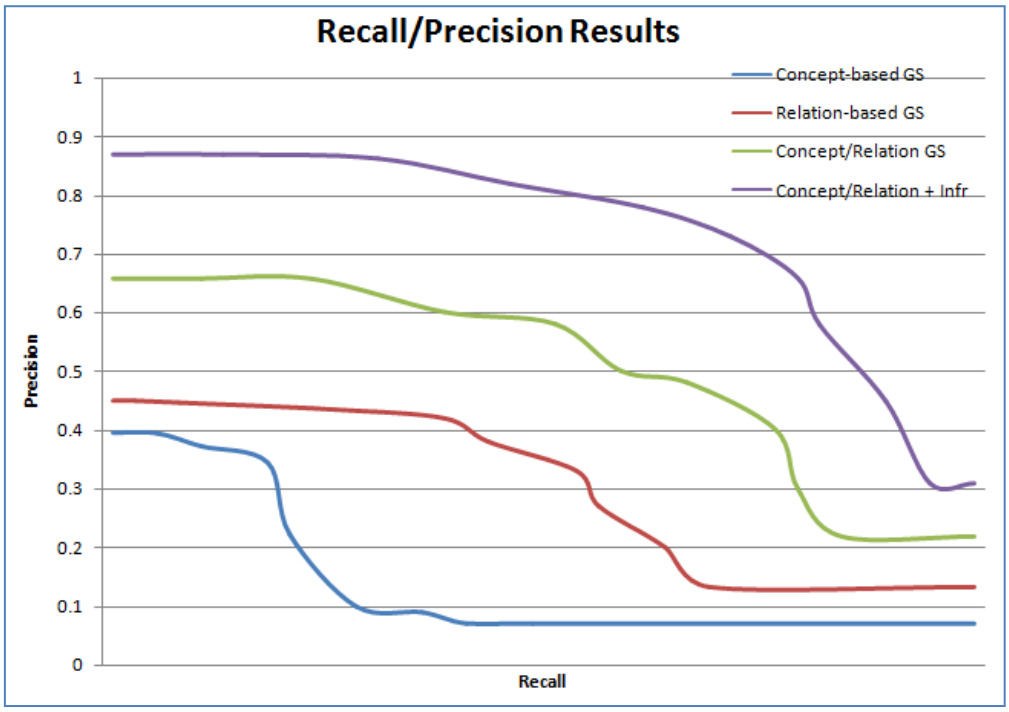



**Figure 6.13: Recall/precision results for the four experimental techniques**

### 6.5.3. Analysis and Discussion

The evaluation results described in Sections 6.5.1 and 6.5.2 revealed that a precision of 87% was achieved. Although searching and retrieving experiments on art/paint conservation has not previously been covered in the literature, related work on experimental data capture and retrieval using: concept-based graph similarities (Baruzzo et al., 2008, Challapalli et al., 2006, Haslhofer et al., 2010, Sanderson and Van de Sompel, 2010, Schmidt et al., 2011); chemistry (Krafft et al., 2010, Pirró et al., 2010, Reid and Edwards, 2009, Altintas et al., 2006, Mcphillips et al., 2006, Hull et al., 2006, Oinn et al., 2004), and natural sciences (Smith et al., 2011, Abidi et al.,

2012, De Roure et al., 2007, Goble et al., 2006, Goble and De Roure, 2007); indicates worse performance than the art conservation results.

For example, the previous approaches that used concept-based graph similarity achieved lower precision of 39.4%. Specific examples of projects that use concept-based graph similarity include:

- E-Dvara (Baruzzo et al., 2008, Challapalli et al., 2006), and Scratchpads (Smith et al., 2011) – these projects use natural language processing techniques (built on Drupal semantic modules) in their recommendation systems.
- Europeana (Haslhofer et al., 2010, Sanderson and Van de Sompel, 2010, Schmidt et al., 2011) – the recommendation system in this project uses the text from the captured annotations.
- Vivo (Krafft et al., 2010), and DOKMS (Pirró et al., 2010) – the Web-based and Desktop-based search interfaces (for research activities and researcher profiles) in this project use Lucene indexing (keyword based indexing).

Furthermore, previous approaches that used relation-based graph similarity achieved higher precision (45%) than the projects included in the above list, but less accuracy than the art conservation experimental data capture. A specific example of a project that uses relation-based graph similarity is:

- POKM (Abidi et al., 2012) – the data recommendation system in this project which aims to discover compatible and relevant services (experiments and specialised experimental workflows) uses services ontology coupled with semantic-based document and content management methods. The final top-k results are determined based on the number of relationships held within the experiments/workflows found by the POKM decision support system.

In the same manner, concept/relation-based graph similarity in systems that use multiple approaches to search and (navigate) experiments (e.g., OurSpaces Virtual Research Environment (Reid and Edwards, 2009), Kepler (Altintas et al., 2006, Mcphillips et al., 2006), Taverna (Hull et al., 2006, Oinn et al., 2004), and myExperiment (De Roure et al., 2007, Goble et al., 2006, Goble and De Roure,

2007)) achieved higher precision (65.7%) than the projects listed above, but less accuracy than the art conservation experimental data capture.

The ontological information, as well as the OWL 2 RL inferencing, added to both the original and recommended graphs (the overall knowledge-base) improved the recommendation results and further boosted the accuracy. The following list provides examples of the inferencing rules that directed the improved results:

- If the given graph includes a sample that was taken from Grumbacher paint tin, and the sample underwent a SEM characterisation activity, and that characterisation identified zinc in the sample, then the recommended graphs will include publications and experiments that include this information, as well as case studies that involve the SEM characterisation of zinc, or Grumbacher paint. This is based on the following inferencing rules included in the OPPRA ontology:
  - Artifact *undergoes* Characterization . Characterization *hasIdentified* Material →Artifact*isComposedOf* Material
  - Artifact*wasSampleSourceOf*Sample . Sample *isComposedOf* Material →Artifact*isComposedOf* Material
  - owl:TransitiveProperty(*isComposedOf*).
- If the given graph includes a spectra file that refers to zinc, and the spectra file was an output of a SEM characterisation activity, then the recommended graphs will include publications and experiments that include this information, as well as case studies that involve zinc samples (and materials) that underwent SEM characterisations. This is based on the following inferencing rules included in the OPPRA ontology:
  - Characterization *outputs* InformationObject . InformationObject*refersTo* Entity → Characterization *characterized* Entity
  - Material *isPartOf* Material . owl:TransitiveProperty(*isPartOf*).

## 6.6. Summary

In this chapter, a Web-based and collaborative experimental workflow system was presented. The system enables researchers (conservators and scientists) to:

- Quickly and easily describe and publish their art/paint conservation data;
- Link experiments to publications (and case studies) via Named Graphs/URIs;

- Find experiments based on keyword, ontology, and graph similarity search.

Using the OPPRA ontology, the Experimental Data Capture platform represents the semantics of experimental data in a form that facilitates re-use and discovery. For example, the system evaluation of the effectiveness of linking experimental data to past publications showed that using specific properties of the OPPRA ontology (concepts, relations and term/relation weightings used for information retrieval recommendation and evaluation tasks), as well as OWL 2 RL inferencing, significantly increased the precision and recall for any given top-k results.

Finally, limitations and future work for the Experimental Data Capture were identified. For example, the Experimental Data Capture component is limited in that:

- It does not support importing data (experiments/sub-experiments) from other content management and experimental workflow systems (e.g., Kepler, Taverna, and myExperiment Virtual Research Environment).In the future, an '*import*' option/functionality may be developed to allow researchers to easily incorporate their previously conducted experiments into the Experimental Data Capture framework, and in turn publish them into the OPPRA-based knowledge-base;
- It does not describe indicative conditional branch statements (i.e., logical operations that act upon OPPRA's entities and govern the experimental workflow/process).Being able to capture and share the workflow patterns associated with experiments would enable greater comparison and re-use of experimental data.

# Chapter 7

# Extracting Structured Data from Past Publications

## 7.1. Introduction

The most common method of publishing new discoveries about art conservation techniques and research has been through traditional full-text publications. Such corpora typically only support searching via metadata (e.g., title, authors or keywords) and full-text. In particular, it is difficult to discover valuable information about the chemical processes, experimental results or preservation treatments associated with the conservation of paintings from a specific genre. This research addresses this problem by focusing on the extraction of structured data (that complies with a pre-defined ontology) from a distributed corpus of publications about painting conservation. The specific extraction method involves a unique combination of NER (using gazetteer-based and ML-based methods) followed by RE (using rule-based and ML-based methods). The resulting structured data is stored in an RDF triple store and a Web-based graphical user interface enables the SPARQL querying, retrieval and display of the search results. The results from applying these techniques to a corpus of publications on art conservation indicate that this approach achieved higher quality precision and recall in extracting NEs and relations from publications, relative to alternative existing approaches.

The 20$^{th}$ Century in Paint project (20thcpaint, 2010b) is a collaborative research project that involves art conservators, chemists, materials scientists and information

scientists working together to build a comprehensive high quality knowledge-base to support more informed decisions by conservators of 20th century paintings.

A major source of legacy information and knowledge on painting conservation is embedded in past publications published via a range of publishers and journals (e.g., JAIC, JSTOR Studies in Conservation, Analytical Chemistry and AICCM Bulletin). The data in these publications is embedded within free-text and is "unstructured data". Such unstructured data is very difficult to query, retrieve, correlate, integrate, analyse or re-use. Moreover, the expansion of publications in this field means it is increasingly infeasible to manually extract all of the data about new discoveries, methods and experiments associated with painting conservation.

The aim of the research described in this chapter is to extract structured knowledge from past publications about art/paint conservation – using semi-automated NER and RE techniques that have been successfully applied to the bioinformatics field. The extracted knowledge is represented in a standardised machine-processable format (compliant with the OPPRA ontology described in Chapter 4) and stored in an RDF triple store – a structured knowledge-base that facilitates integration and correlation with other publications, experimental data and external databases (e.g., on paint chemistry) and fast, ontology-based searching and browsing.

The OPPRA ontology underpins both the NER and RE steps. For example, OPPRA defines entities such as: paint names (e.g., "Winsor & Newton"), paint types (e.g., "acrylic"), pigments (e.g., "chrome yellow"), deterioration mechanisms (e.g., "blistering"), experimental procedures (e.g., "artificial aging"), characterisation techniques (e.g., "Fourier Transform Infrared Spectroscopy" or "FTIR"), conservation methods (e.g., "cleaning") and materials (e.g., "solvent"). OPPRA also defines the relationships between detected NEs, such as: "Winsor & Newton *manufactured* Artists' Acrylic Cadmium Yellow", "CitricAcid *hasChemicalFormula* C6H8O7", and "Assessment001 *hasIdentified* Blistering".

This proposal is to use the OPPRA ontology as the underlying gazetteer, combined with a rule and supervised ML-based approach to extract meaningful statements

from free-text publications that express facts or accepted knowledge associated with paint conservation. The specific objectives of this work are to:

- Acquire a corpus of publications about art/paint conservation;
- Develop, apply and optimise a NER service (using OPPRA as the gazetteer);
- Develop, apply and optimise a RE service;
- Develop a graphical user interface to enable users to interactively review, correct and refine extracted triples;
- Save the structured data in an OPPRA-based RDF triple store;
- Provide a search, browse and query interface over the triple store.

## 7.2. Related Work

Information extraction methods applied to extract structured data from publications have been studied extensively in the past in fields that include biology (Barbosa-Silva et al., 2010, Corney et al., 2004, Meurs et al., 2011), chemistry (Na et al., 2010, Yamashita et al., 2011), biomedicine (Bundschus et al., 2008, Crowley et al., 2010, Pestian et al., 2007), text summarisation (Lin and Hovy, 2000) and e-learning (Monachesi et al., 2009). A review of these previous efforts is provided in this section.

Generally, structured data extraction can be broken down into two steps, namely, NER and RE. Previous approaches to the NER task can be divided into two main categories:

- Gazetteer-based NER is the process of detecting NEs in text documents based on predefined lists of synonyms (or phrases). Examples of implementations that use gazetteers include Maynard et al. (2009) and Navigli and Velardi (2008).
- ML-based NER relies on trained classifiers (supervised or unsupervised) to detect NEs in text documents, using a set of features extracted from the given text. For example, Hovy et al. (2009) and Mintz et al. (2009) detected NEs using tokens, Part-of-Speech – PoS and morphological analyses of textual sentences.

Past approaches to the RE task have employed: syntactical analysis and shallow parsing. Syntactical analysis is the process of analysing strings of symbols (e.g., sentences) according to the rules of their formal language (e.g., English). Examples

of syntactical analysis approaches include: Jurij et al. (2005) used a Support Vector Machine – SVM on the deep syntactic analysis produced by NLPWin (Dale et al., 2000) to augment a detailed graph for mentioned entities which in turn were summarised to produce summary sentences of the given text; Pandit and Honavar (2010) trained the Stanford Parser's Dependencies with hand-crafted rules for ML in order to extract RDF from complex relationships occurring in text sentences. The further discussion below is based on shallow parsing (analysis that identifies nouns, noun groups, verbs, verb groups) in the related literature. This is more appropriate than syntactical analysis because the focus of the present study was on extracting the simplified core knowledge that conforms to a defined ontology.

The RE shallow parsing methods can be categorised into two alternate approaches:

- Rule-based RE depends on the definition of rules surrounding noun and verb phrases to determine if a relation holds between the NEs. For example, Genereux and Niccolucci (2006) and Maynard et al. (2009) used sets of rules that included windows of noun and verb phrases to extract relations (e.g., *subClassOf* and *sameAs*) and properties (e.g., *Birds have feathers*) from text documents.
- ML-based RE determines the presence of relationships between NEs using trained classifiers with features such as tokens, PoS, morphological analysis, and NEs (nouns, verbs and phrases). For example, Mintz et al. (2009) built unsupervised classifiers by filtering 102 relationships present in Freebase. Byrne and Klein (2010) used SVM combined with verbs and their surrounding entities (e.g., *actors, dates* and *locations*) to extract *find*, *visit* and *move* events from archaeological texts.

Such approaches have also been successfully applied in the biomedical domain. Embarek and Ferret (2008) attempted to extract four relations (*detect, treats, sign of, cures*) between five types of medical entities, based on patterns which were automatically built using an edit distance between sentences and a multi-level phrase matching algorithm. Schneider et al. (2009) used syntactic patterns over parsed text, surface patterns and automatically learned ‘transparent words’ to detect protein-protein interactions. Roberts et al. (2008) used ML techniques to detect

which semantic relationship linked two occurrences of medical entities in a sentence. Finally, Grouin et al. (2010) used a hybrid approach to extract relations from clinical texts: a pattern-based method is applied first, then for sentences where no relation was found, a ML method is automatically applied.

Furthermore, similar approaches (ontology-based GATE gazetteer and relation extraction using pre-defined rules, and machine-learning) have been applied with mixed success to publications in the bioinformatics field (Barbosa-Silva et al., 2010, Bundschus et al., 2008, Corney et al., 2004, Crowley et al., 2010, Meurs et al., 2011, Na et al., 2010, Pestian et al., 2007, Yamashita et al., 2011). Further examples are provided in the following paragraphs.

The oreChem ChemXSeer project (Na et al., 2010) combined the OAI-ORE and OreChem ontologies to describe chemicals, chemical processes and experiments to extract experiment information from chemistry publications (e.g., experimenter, reaction, input, and output) which were represented as OAI-ORE compound objects. This approach used SVM by considering the number of keywords (and NEs predicted in an earlier stage) in each paragraph.

In CULTURA, both IBM LanguageWare (IBM, 2013) and GATE (Cunningham et al., 2002) are used to perform entity extraction. CULTURA enables experts to visualise the entity graph (exported after the entity extraction process occurs) using the PreMapper tool developed by Commetric (Commetric, 2013). PreMapper enables curators of data to add, delete, merge, disambiguate and edit entities using a graphical user interface. The transparency of this tool helps experts to identify errors in the entity graph, and allows them to manually correct the output of the automatic entity extraction process.

A pipeline called txt2rdf (Byrne, 2009) was proposed to augment the cultural heritage structured data with "facts" automatically extracted from free text. The pipeline takes in a plain text at one end, and outputs RDF triples combined with related Semantic Web data. The pipeline shows an accuracy of 76% for NER and RE, and it has been shown to produce an integrated RDF graph structure that can answer queries for information that was impossible to retrieve previously.

A review of the literature, however, has revealed that approaches that focus on structured knowledge extraction from full text publications about art/paint conservation (to help conservators identify the cause of paint deterioration and the optimum treatment or environment to remove or prevent further deterioration) do not exist.

## 7.3. Methodology for Extracting Structured Knowledge

This section describes the four main steps involved in developing structured knowledge extraction system:

- Development of the OPPRA ontology;
- Publication collection and manual annotation;
- Implementation of the NER tool;
- Implementation of the RE tool.

### 7.3.1. The Ontology of Paintings and PReservation of Art – OPPRA

One of the first tasks in the 20th Century in Paint project is the development of the OPPRA ontology (Hunter and Odat, 2011) to capture the semantics of paints and painting preservation descriptions, and to support provenance-related queries. The OPPRA ontology is used to build the gazetteer that identifies NEs that occur within the text of past publications (e.g., *Sidney Nolan, Melbourne, the 20th century, darkening, FTIR* and *Ripolin*) and to map them to their corresponding classes (e.g., *Artist, Place, Date, Condition Change, Microscopic Technique* and *Paint*).

The OPPRA ontology is also used by the rule-based and ML-based RE modules to define the relationships between these NEs. Figure 7.1 highlights the main classes and properties defined in the OPPRA ontology (e.g., *undergoes*(Material, Event), *carriedOutBy*(Activity, Actor), *hasCondition*(Painting, ConditionState), *removes*(Solvent, Varnish), *hasChemicalStructure*(Material, ChemicalStructure), and *hasArtist*(Painting, Artist)). Specific details on the complete OPPRA ontology are published (20thcpaint, 2010a), and were given in Chapter 4. An OWL representation of the OPPRA ontology can also be viewed online (20thcpaint, 2010a).

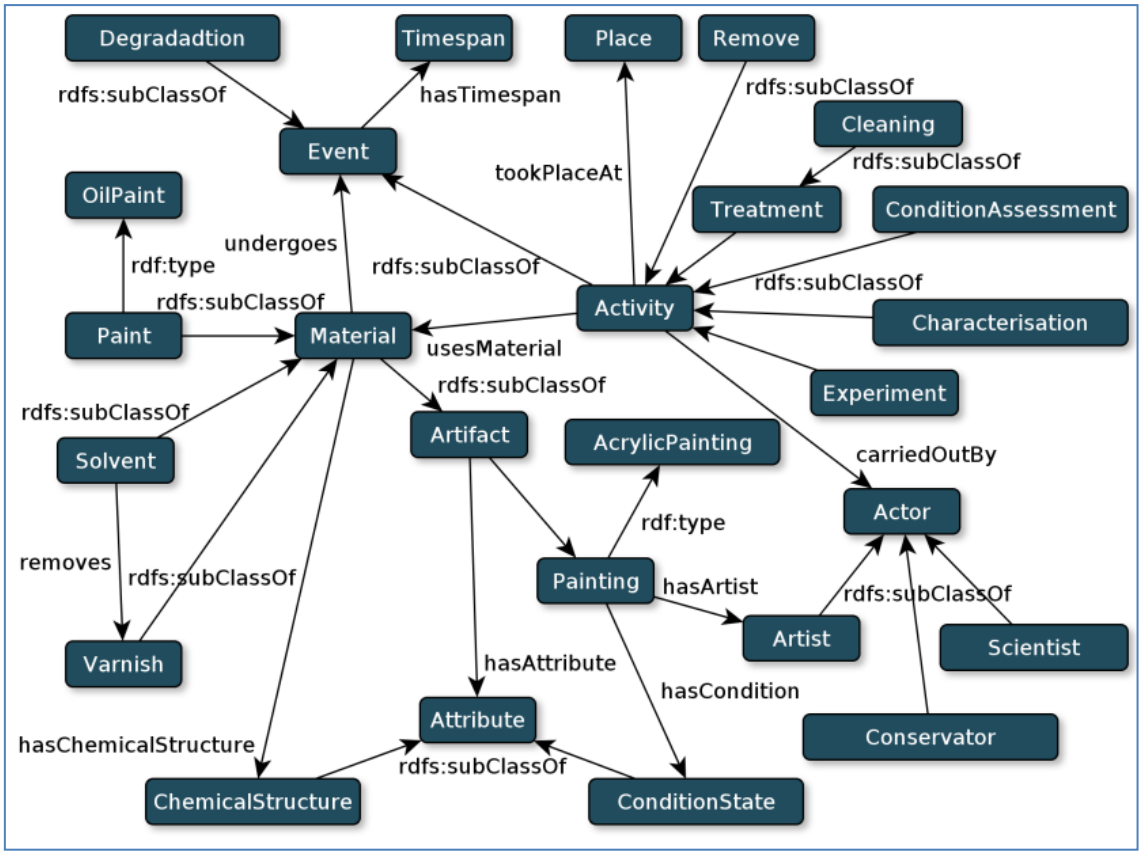



**Figure 7.1: OPPRA's main classes and properties**

### 7.3.2. Publication Collection and Manual Annotation

The next step involved identifying a corpus of relevant publications about paint conservation. This corpus is used to train and evaluate the ML algorithms – ambiguity resolution and RE. Altogether, 100 publications from 15 different journals were identified with input from the art conservators and material scientists working on the 20$^{th}$ Century Paint project. The publications/journals were chosen on the basis of both quality (impact factor and citations) and relevance. The chosen journals included the top ranked journals in the field: the JAIC (COOL, 2002), JSTOR Studies in Conservation (JSTOR, 2000), Analytical Chemistry (ACS, 2011), and AICCM Bulletin (AICCM, 1973). Copyright issues were not a factor in this process because the documents were indexed/annotated for search/querying purposes only – links to the full-text of the document were provided if the user wished to proceed to read the document (based on his/her access rights in the publisher's Website).

After the corpus of relevant publications on paint conservation was selected, 15 articles were selected for training and testing purposes. These publications were

manually tagged to generate RDF descriptions that conformed with the OPPRA ontology. For example, consider the publication:

'*MONICO, L., VAN DER SNICKT, G., JANSSENS, K., DE NOLF, W., MILIANI, C., DIK, J., RADEPONT, M., HENDRIKS, E., GELDOF, M. & COTTE, M. 2011. Degradation Process of Lead Chromate in Paintings by Vincent van Gogh Studied by Means of Synchrotron X-ray Spectromicroscopy and Related Methods. 2. Original Paint Layer Samples. Analytical Chemistry, 83, 1224-1231.*'

Manual tagging of the textual content in this paper generated RDF instance data corresponding to the structured/modelled information shown in Figure 7.2.

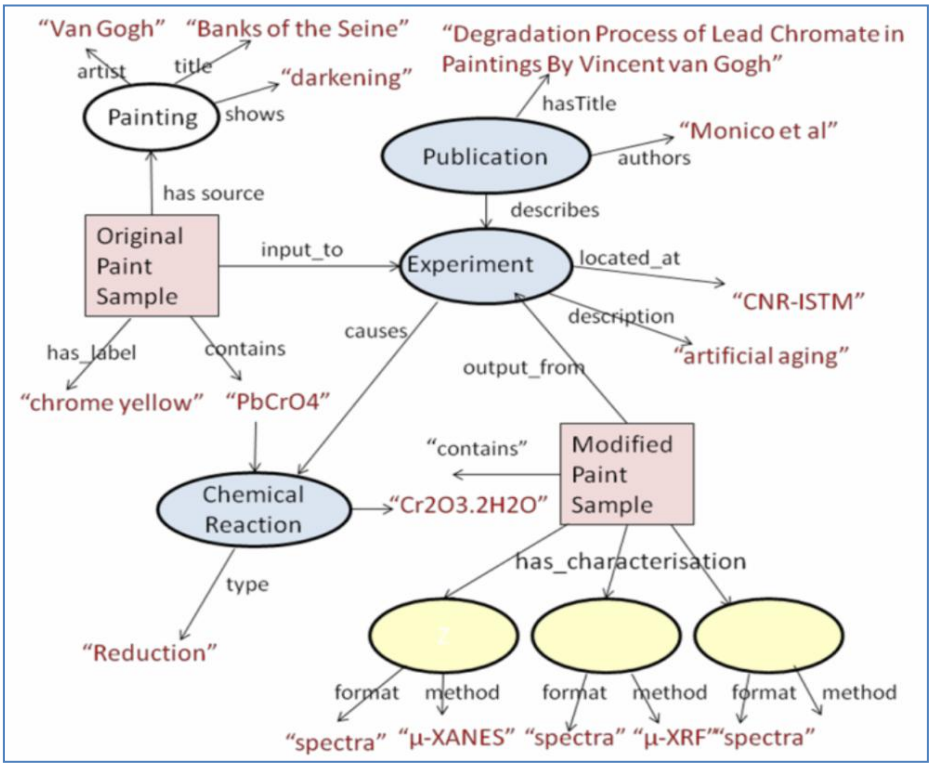



**Figure 7.2: Structured data extracted from an example publication**

The manual annotation of text documents was done by the author using the *text2triples* software. Further details on the text annotation task are provided in Sections 7.4.1, and 7.4.2. Initially, the manual annotation task on 15 articles was performed. The annotations were then verified by the 20th Century in Paint team members. There were no inter-annotator agreement tests. Text documents typically

comprised an average of 160 sentences, and 1030 annotations. Further details on the document statistics are given in Section 7.5.1 (the experimental results section).

The average time taken to manually annotate a text document was 3-5 hours. The manual annotation of the training corpus was very time-consuming because it was important to ensure that it was high quality for training and testing the NER system. Publications that had firstly undergone automatic annotation using the trained NER system only required an additional 10-30 minutes of manual annotation to check/correct the annotation results.

### 7.3.3. Named Entities Recognition – NER

The approach to NER is to combine gazetteer lists complemented with a classifier to identify NEs (e.g., *Sidney Nolan, Melbourne, the 20th century, darkening, FTIR* and *Ripolin*) mentioned in text publications, and to map these NEs to their proper URIs (classes) in the OPPRA ontology (e.g., *Artist, Place, Date, Condition Change, Microscopic Technique* and *Paint*). This section describes the processes involved in the NER step.

*OPPRA-based Gazetteer*

An OPPRA-based gazetteer is developed by extending the GATE OntoRoot Gazetteer (Danica et al., 2008). The extension gathers synonyms in the OPPRA ontology as a list of terms, matches these terms against the tokens' roots in the text documents and tags these results with their appropriate URIs in the ontology.

Compared to the GATE OntoRoot Gazetteer, this approach adds four new features. Firstly, the OPPRA-based gazetteer performs lookups on synonyms rather than class URIs. This enables the detection of exact matches of texts such as *spaces*, *special characters* and *long titles or sentences*. Secondly, fractions of texts (e.g., *darkened* as a verb or adjective) were enabled to have multiple URIs (e.g., *oppra:Darkening* / *oppra:Darkened*). Although this feature leads to ambiguities, it improves recall and the ambiguities are resolved later via the classifier.

The third modification made is to introduce the ability to clean overlapping lookups. This enables smooth feature extraction as well as the assignment of a contextual

meaning to the terms. The following list provides some examples (texts within "*[]*" and "<>" denote tagged lookups):

- **WithinSpanOf**: *<The [recovery] of [color] in [scorched] [oil paint] [films]>→oppra:publication*;
- **Overlaps**: *<long wavelength [ultraviolet> radiation]→* oppra*:LongWavelengthUltraviolet*;
- **Exact boundaries**: *During <[<bleaching>]> the paint samples became only slightly warm→ambiguities (oppra:Bleaching, oppra:Bleach and oppra:Bleached)*.

Finally, the OPPRA-based gazetteer caches only synonyms in memory (other properties are obtained at runtime). During the testing of the NER task, an out of memory exception occurred when loading the entire ontology into the Gate OntoRoot Gazetteer. This was due to the large amount of data that is stored in memory (e.g., complete URIs, label data/annotation properties, subclass and equivalent object properties). Storing only synonyms in memory (and obtaining the other properties as needed) resolves the out of memory exceptions that are inherent to large ontologies.

Figure 7.3 shows an example of a document tagged using the OPPRA-based gazetteer. The terms and instances highlighted in green were mapped to the OPPRA classes with a certainty of 100% (no ambiguities detected), the red highlighted terms were mapped to OPPRA classes with ambiguities resolved by the NER classifier (via suggestions to be corrected by the user in a later stage), and the yellow highlighted items were terms of interest that would potentially require users to select and add new terms to OPPRA.

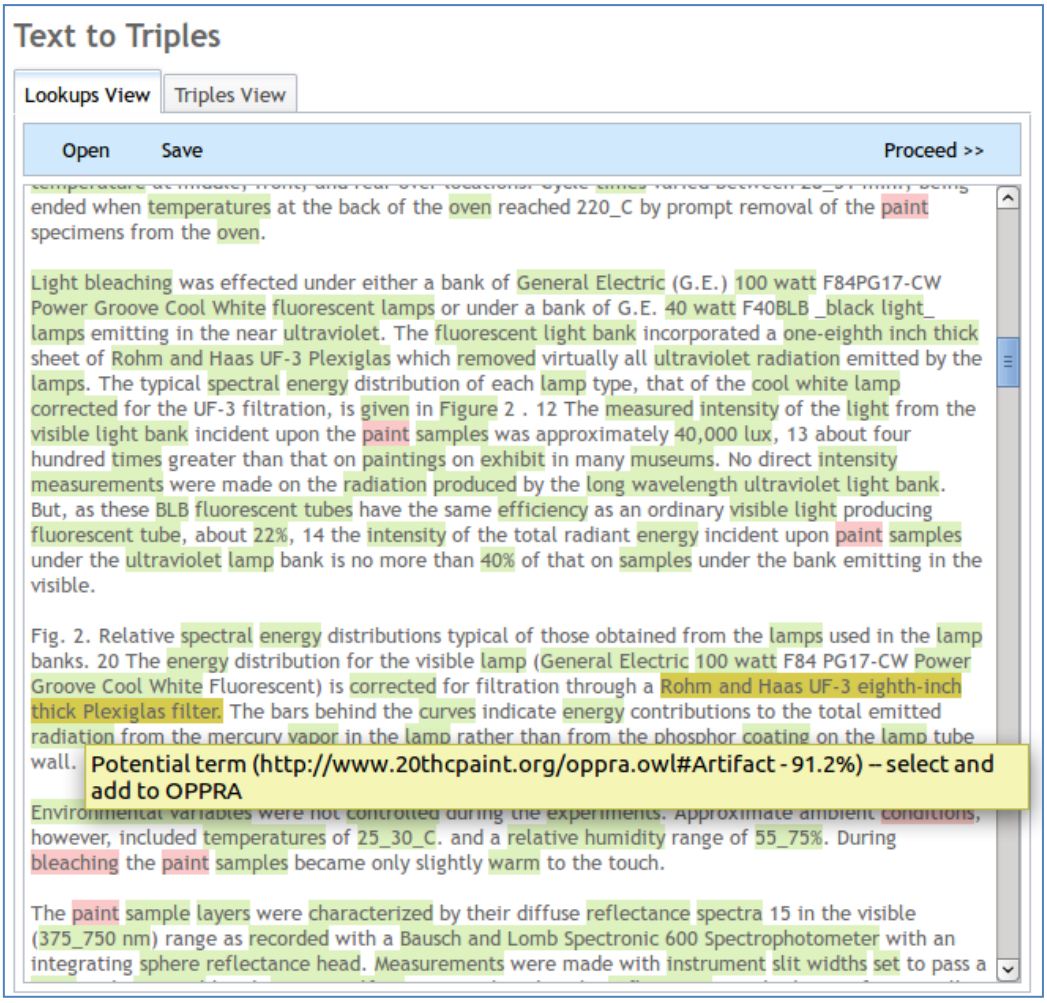



**Figure 7.3: OPPRA-based gazetteer executed on GATE**

*NER Classifier*

After applying the OPPRA-based gazetteer, finding entities that are not present in the OPPRA ontology, and to resolving ambiguities are needed. This problem is treated as a sequential labelling task. The MALLET ML toolkit (McCallum, 2002) was chosen because it provides implementations of widely-used sequence algorithms including Hidden Markov Models – HMMs and linear chain Conditional Random Fields – CRFs. Moreover, MALLET CRFs have been previously applied to text processing in a variety of domains, including bioinformatics, to perform NER, Dependency Parsing and Co-reference Resolution (Kudo et al., 2004, McCallum and Li, 2003, Qi et al., 2005).

The NER classifier was trained using a set of features and labels (predictions) extracted from sentences (MALLET instances) within the manually annotated corpus

(Section 7.2.2). Each word in a sentence was regarded as a token and each token was associated with a set of features and a label.

The extracted features and the reason for their usage, comprise:

- Root features: This provides a better coverage since words are compared based on their roots (e.g., darkening = darkened = darken);
- PoS features: This enables context-based detection of a specific word's meaning – i.e., based on the word's usage within a sentence (e.g., ambiguity resolution);
- Orthographical features (and prefix and suffix features): This enables the shape of words to be identified for a better coverage, and context-based detection of a word's meaning (e.g., INTRODUCTION (uppercase), darkening (lowercase), Nolan (upper-initial), and the 's' suffix in 'darken**s**').

The above features are produced from the following GATE plugins:

- ANNIE English Tokenizer: this plugin splits the text into very simple tokens such as numbers, punctuation and words of different types. Orthographical features "*orth*" distinguish between words in upper-initial, lowercase and uppercase (e.g., *upperInitial, lowercase* and *allCaps*);
- ANNIE POS Tagger: this plugin produces a PoS tag (GATE, 2012)as an annotation on each word or symbol. Examples of the PoS features used include: *NN* (noun), *VB* (verb), *JJ* (adjective), *DT* (determiner), *IN* (preposition or subordinating conjunction), *CC* (coordinating conjunction), *RB* (particle), *TO* (literal '*to*'), *NNS* (noun – plural), *VBG* (verb – gerund or present participle), *VBZ* (verb – 3rd person singular present), *VBN* (verb – past participle) and *RBR* (adverb – comparative adverbs);
- GATE Morphological Analyser: the morphological analyser takes as input a tokenised document and considers the document's tokens and their PoS tags (one at a time) to identify their lemmas (*root*) and prefixes and postfixes (*affix*).

Each label (prediction) with a form of *B-Lookup, I-Lookup* or *0* indicates not only the OPPRA class that the NE belongs to, but also the location of the token within the NE. Using this notation, Lookup was the OPPRA class label; B and I were the location labels for the beginning of an entity and the inside of an entity, respectively; and 0 indicated that a token was not part of an NE.

Table 7.1 illustrates an example of a training instance for the sentence '*In separate experiments involving longer light exposure periods, however, flake white paint has also been found to exhibit this bleaching behavior.'* Column 1 shows the features extracted from GATE plugins (above) for each token in the sentence; and column 2 shows the labels/predictions assigned for the NER classifier.

**Table 7.1: MALLET training/testing features and label lookups for the NER classifier**

| Tokens' Features | Lookups |
|---|---|
| in IN upperInitial | 0 |
| separate JJ lowercase | 0 |
| experiment NNS s lowercase | B-Experiment |
| involve VBG ing lowercase | 0 |
| longer RBR lowercase | 0 |
| light JJ lowercase | B-Description |
| exposure NN lowercase | I-Description |
| period NNS s lowercase | 0 |
| , , | 0 |
| however RB lowercase | 0 |
| , , | 0 |
| flake NN lowercase | B-FlakeWhite |

| Tokens' Features | Lookups |
|---|---|
| white JJ lowercase | I-FlakeWhite |
| paint NN lowercase | B-Paint |
| have VBZ s lowercase | 0 |
| also RB lowercase | 0 |
| be VBN en lowercase | 0 |
| find VBN ed lowercase | 0 |
| to TO lowercase | 0 |
| exhibit VB lowercase | 0 |
| this DT lowercase | 0 |
| bleach VBG ing lowercase | B-Bleaching |
| behavior NN lowercase | 0 |
| . . | 0 |

### 7.3.4. Relation Extraction – RE

The following two sub-sections provide details on the two processes used to extract relationships between the entity lookups (rule-based RE and ML-based RE). Examples of relations that were extracted include:

- *hasAttribute*(Artifact, Attribute) and its sub-properties (e.g., *hasStructure*, *hasComposition* and *hasCondition)*;
- *carriedOutBy*(Activity, Actor) and its sub-properties (e.g., *performedPainting* (PaintingProcess, Artist) and *transferredTitle*(Acquisition, Actor));
- *isComposedOf*(Artifact, Material) and its sub-properties (e.g., *hasSupportType*(Painting, Support), *hasFrameType*(Painting, Frame) and *containsMaterial*(Material, Material));
- *rdfs:type* including instant assignment (e.g., SidneyNolan *rdfs:type* Artist), ID/URL creator (e.g., Sample001 *hasOrganizationalID* "pbcr-Y" and Publication001 *hasURL* "http://example.com") and description detector (e.g., Experiment *hasDescription* "light exposure");
- *tookPlaceAt*(Activity, Place);
- *hasDate*(Event, Date).

*Rule-based RE*

The rule-based RE task identifies relations that exist between NEs within noun and verb phrases. It also bootstraps the ML predictions (described in the next section – ML-based RE) by decreasing the number of tokens in each training/testing instance. As an example, the rule-based RE takes the phrase *'flake white paint*', extracts features from it, and compares the features with a set of rules to generate the RDF triple *containsMaterial(oppra:Paint, oppra:FlakeWhite)*.

Rules define allowable sequences between the PoS (e.g., *DT* (determiner), *JJ* (adjective), *CC* (coordinating conjunction)) and lookup types (classes and instances) in the gazetteer/OPPRA ontology. The rules were derived using the training corpus of 15 documents. All noun and verb phrases where extracted, then tokens' PoS and lookup types within these phrases were selected. The appropriate relationships (e.g., *id, subject, has Attribute*) were recorded manually. To test how robust the rules are with respect to minor changes in phrasing and a wider range of documents, the coverage of the recorded rules was tested. Specific details on the experimental setup and results can be found in Section 7.5.2. The experimental results used 3370 noun and verb phrases (from the corpus of 5 testing documents), and an F-Measure of 98.36% was found. The high coverage is due to the large amount of, but small length of, training instances (noun/verb phrases).

Table 7.2 provides examples of rules applied to noun phrases and their corresponding outputs (triples). For example, given the phrase "*the desired <pigment> and <medium><composition>*", the rule-based RE constructs the sequence "*the* DT / *desired* JJ / *pigment* Class / *and* CC / *medium* Class / *composition* Class" to produce the triples *isAttributeOf(Composition, Pigment)* and *isAttributeOf(Composition, Medium)*.

**Table 7.2: Examples of rules and triples**

| Rules | Examples | Triples |
|---|---|---|
| DT JJ Class CC Class Class | the desired <pigment> and <medium><composition> | Composition isAttributeOf Pigment<br>Composition isAttributeOf Medium |
| Instance Class | <poppyseed><oil> | Poppyseed (subject – instance of Oil) |
| Instance Class CC Instance Class | <carbon dioxide><gas> and <water><vapor> | CarbonDioxide (subject)<br>Water (subject) |
| Class Class Class CC Class | <scorched><paint><medium> and <pigment> | Paint containsMaterial Medium<br>Paint containsMaterial Pigment<br>Paint hasAttribute Scorched |
| Instance Instance Class | <white><linseed oil><house paints> | HousePaint containsMaterial LinseedOil<br>HousePaint hasAttribute White |
| DT Instance Class Instance | the <yellow><paint><cadmium yellow deep> | Paint containsMaterial CadmiumYellowDeep<br>Paint hasAttribute Yellow |

*ML-based RE*

The MALLET CRF implementation for the ML-based RE classification procedure was used because it provides implementations of widely-used sequence algorithms as explained in relation to the NER classifier above. In this phase, a classifier for each relation present in the OPPRA ontology is constructed. For example, if the relation "*undergoes(Artifact, Event)*" is to be extracted, then a classifier (named *undergoes*) needs to be constructed. The steps involved in constructing this construction process are as follows:

- The output from the rule-based RE step above was taken. This replaces each phrase with its subject lookup/NE. For example, the sentence '*In separate experiments involving longer light exposure periods, however, flake white paint has also been found to exhibit this bleaching behavior*' would be reduced after the rule-based RE task to new lookups/NEs (represented in *bold* text): *In **experiments** involving light exposure, however, **flake white** 'has also been' found 'to exhibit' this **bleaching** behavior.*
- A window of lookups/NEs (between 2 and *n*) was used to form MALLET instances. For example, the above output from the rule-based RE would generate the following training/testing instances:
  - ***experiments** involving light exposure, however, **flake white**;*
  - ***experiments** involving light exposure, however, **flake white** 'has also been' found 'to exhibit' this **bleaching**;*
  - ***flake white** 'has also been' found 'to exhibit' this **bleaching**.*
- Next, the features and label prediction ('*0*' indicating that there is no relation; '*1*' indicating that there is a forward relation and '*-1*' indicating that there is a

backward relation between the NEs within the specified window) for each token were incorporated. The following corresponds to the MALLET instances for the example sentence:

- **experiments** experiment NNS s lowercase B-Experiment / **involving** involve VBG ing lowercase / **light** light JJ lowercase / **exposure** exposure NN lowercase / **,** , , / **however** however RB lowercase / **,** , , / **flake** flake NN lowercase B-FlakeWhite / **white** white JJ lowercase I-FlakeWhite → -1 (*undergoes(FlakeWhite, Experiment)*);
- **experiments** experiment NNS s lowercase B-Experiment / … → 0 (no undergoes relation);
- **flake** flake NN lowercase B-FlakeWhite / **white** white JJ lowercase I-FlakeWhite / **been** be VBN en lowercase / **found** find VBN ed lowercase / **exhibit** exhibit VB lowercase / **this** this DT lowercase / **bleaching** bleach VBG ing lowercase B-Bleaching → 1 (*undergoes(FlakeWhite, Bleaching)*).

- Finally, a set of new training instances using the semantic parent of each lookup/NE (in the gazetteer/OPPRA ontology) was generated. This step ensured that the corpus (15 publications) was large enough to produce accurate results from the ML-based RE task. For example, *FlakeWhite* and *Bleaching* in the training instance '***flake white*** *has also been found to exhibit this* ***bleaching***' can be replaced with their semantic parents (*Pigment, Material, Artifact and Entity*) and (*Activity, Event and Entity*) to produce 19 more training instances, including:
  - ***pigment*** *has also been found to exhibit this* ***bleaching*** *→ 1;*
  - ***material*** *has also been found to exhibit this* ***bleaching*** *→ 1;*
  - ***artifact*** *has also been found to exhibit this* ***bleaching*** *→ 1;*
  - ***entity*** *has also been found to exhibit this* ***bleaching*** *→ 1;*
  - ***flake white*** *has also been found to exhibit this* ***activity*** *→ 1;*
  - ***pigment*** *has also been found to exhibit this* ***activity*** *→ 1;*
  - ***material*** *has also been found to exhibit this* ***activity*** *→ 1.*

## 7.4. System Implementation and User Interface

### 7.4.1. System Architecture

A Web-based framework (*text2triples*) is developed to automatically extract structured data from past publications. The framework is implemented using a

combination of Web 2.0, Apache Tomcat, Java implementations of GATE (Cunningham et al., 2011), MALLET (McCallum, 2002) and OpenRDF Sesame Triple Store (Aduna, 1997). HTML 5, JavaScript, Dojo and InfoVis are the underlying technologies in the client side chosen to ensure dynamic Web interfaces and highly responsive interactivity. Figure 7.4 shows the overall architecture of the *text2triples* system.

In the server side, the OPPRA ontology and its RDF instances are stored in an OpenRDF repository – Sesame Triple Store which provides access to OWL/RDF metadata via SPARQL (W3C, 2008) queries. The server also includes a MALLET Java implementation that is called by the GATE pipeline to perform RE and ambiguity resolution based on the given models (MALLET-trained classifiers for each manually tagged concept, NE and relation).

The Java implementation of GATE reads text documents that are uploaded into the system and gets ready to prepare and execute a pipeline based on the user's requests – implemented as the '*QueryGate*' processing resource. Text documents are then pre-processed via an XML character escaping resource that converts all illegal HTML characters to '_' to avoid displaying unreadable characters in the user interface. RegEx Sentence Splitter, English Tokenizer, POS Tagger and VP Chunker from the 'ANNIE' plugin (Cunningham et al., 2002) are used to markup sentences, tokens, parts-of-speeches and verb phrases, respectively. The GATE Morphological Analyser (Aswani and Gaizauskas, 2010) from the 'Tools' plugin is used to add morphological features to tokens. The Noun Phrase Chunker (Munpex, 2012) calls the ANNIE JAPE Transducer to find segments of noun phrases and their subjects within text.

The features extraction in this implementation include: 1) the OPPRA-based gazetteer that extends the OntoRoot Gazetteer (Danica et al., 2008) class to perform NE recognition; 2) the Triples Creator that calls the ML module and responds to users' modifications for all transactions on the extracted triples; and 3) the RDF/XML/JSON Builder that transforms lookups and triples within GATE documents to their corresponding format – for Sesame and UI interactions.

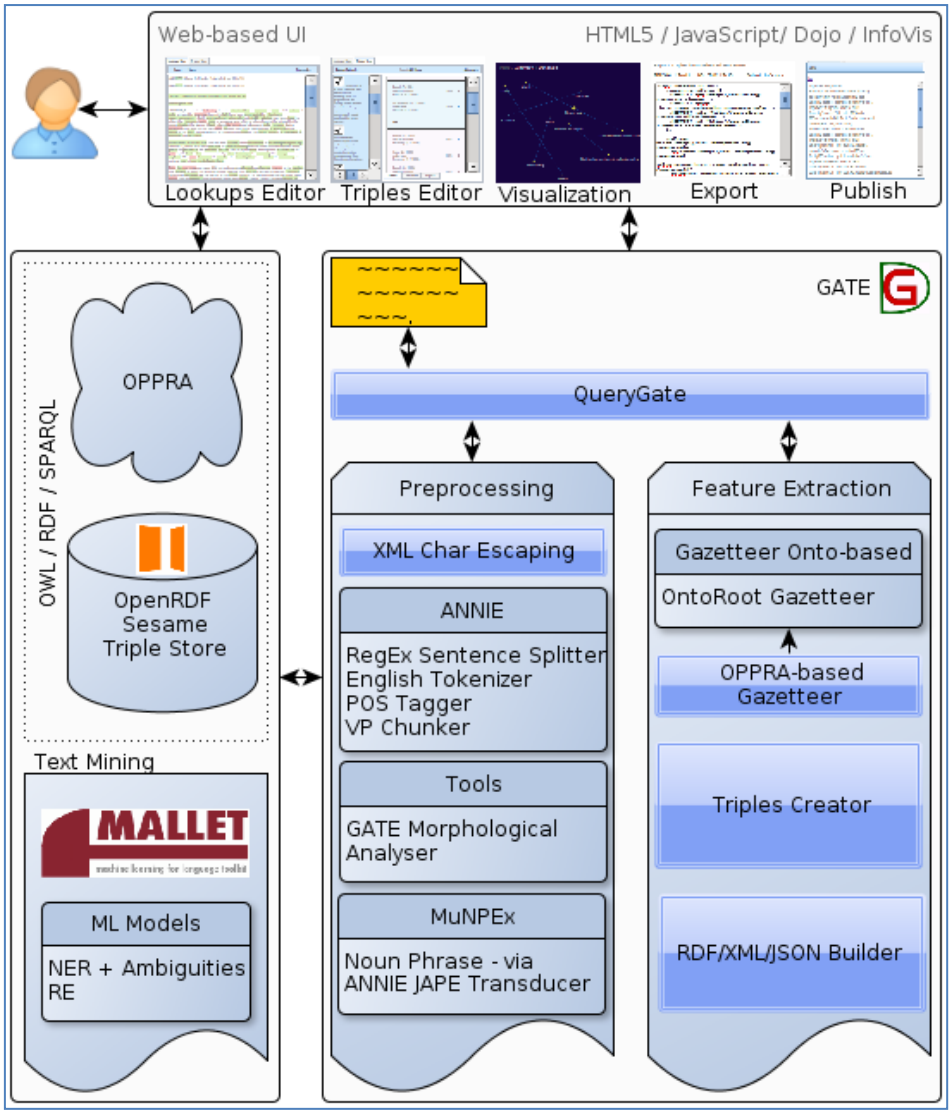



**Figure 7.4: Overview of *text2triples* system architecture**

### 7.4.2. User Interface

A Web-based semi-automatic user interface was developed to enable conservators and scientists to upload text documents, save the generated GATE documents and automatically tag and define relationships between entities within text. The user interface enables users to correct erroneous predictions, save, visualise, export and publish results. The user interface and functionalities are accessible via the Web portal (20thcpaint, 2012b).

As illustrated in Figure 7.3 (Section 7.2.3 above), the lookups editor automatically tags terms and NEs within text using the OPPRA-based gazetteer – the green, red and yellow colours indicate a certainty of 100%, ambiguity detection, and potential terms that may be added to OPPRA, respectively. A *Text Annotation* editor enables

users to add new terms to OPPRA. The editor uses a WordNet service to find synonyms of the selected term, allowing users to select/deselect synonyms and choose the class where the new term is to be placed (in OPPRA). The user can also specify a Web link to the term (e.g., Wikipedia article) which can be used for future visualisations (e.g., *Wikipedia* illustration of a NE *onHover*).

Figure 7.5 illustrates the user interface that displays the extracted triples sequentially in the order in which they occur in the text. The panel on the left shows the sentences from which the triples were extracted. Selecting the Editor tab at the bottom of the Webpage displays the extracted triples in the panels on the RHS and enables users to edit/add/delete them. Users are able to select and deselect sentences from the text document, call the automatic extraction method to find the triples within the selected sentences, and to modify (add and delete) the triples found by the classifier.

The user interface also provides a *Visualisation* tab at the bottom of the page. This displays entities (as stars where the size indicates the number of times that the entity occurs) and relations within the selected sentences as arcs between the nodes. Users can choose to view either a summary of the complete RDF graph corresponding to a textual document or the complete detailed RDF graph.

Once the user is satisfied with the correctness of the RDF triples, the *Export* tab at the bottom of the Webpage enables users to transform the triples to RDF, Turtle, N3, TriX and TriG formats, and to publish the RDF data to the OPPRA-based knowledge-base. Figure 7.6 illustrates screenshots of the visualisation and exporting/publishing interfaces.

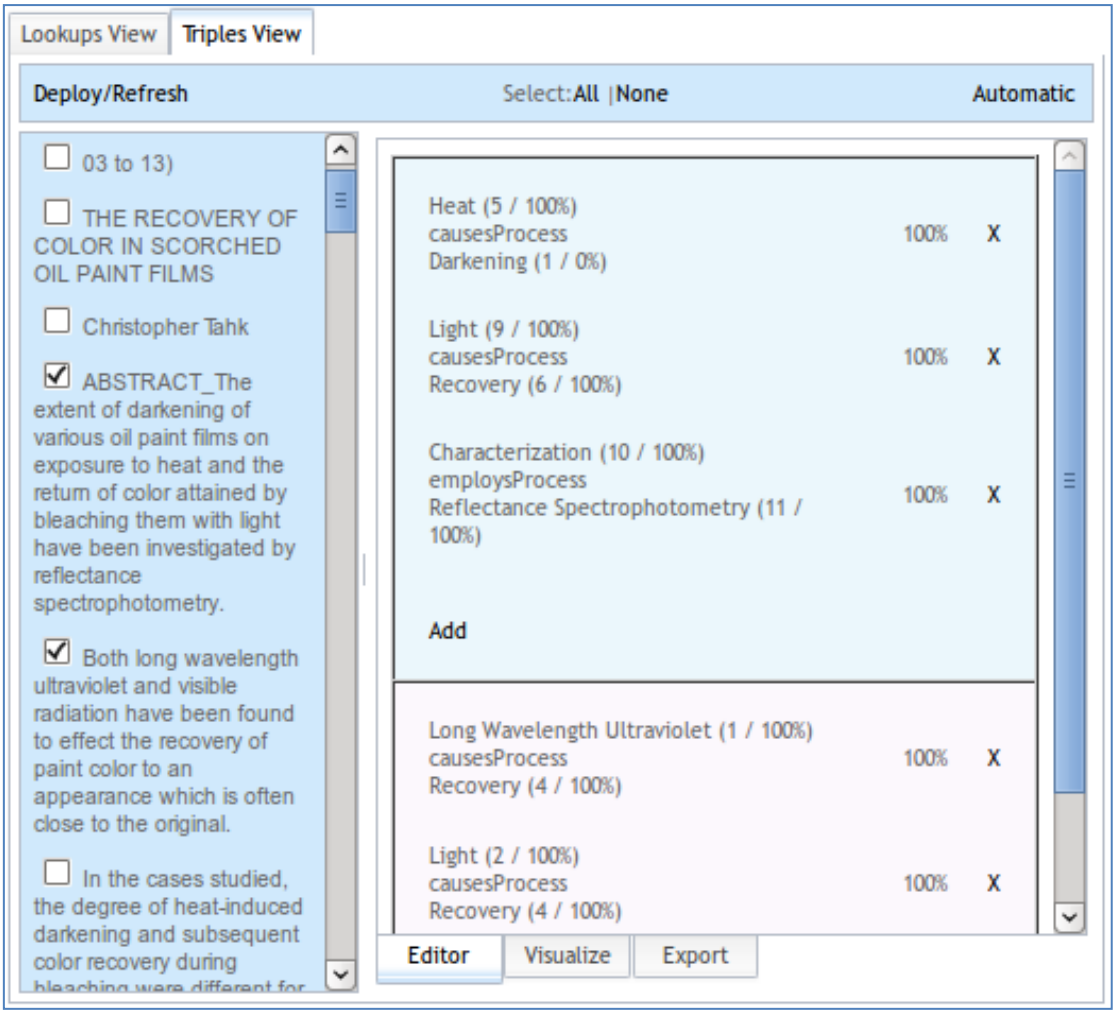



**Figure 7.5: User interface allowing users to select a sentence and edit the automatically extracted triples**

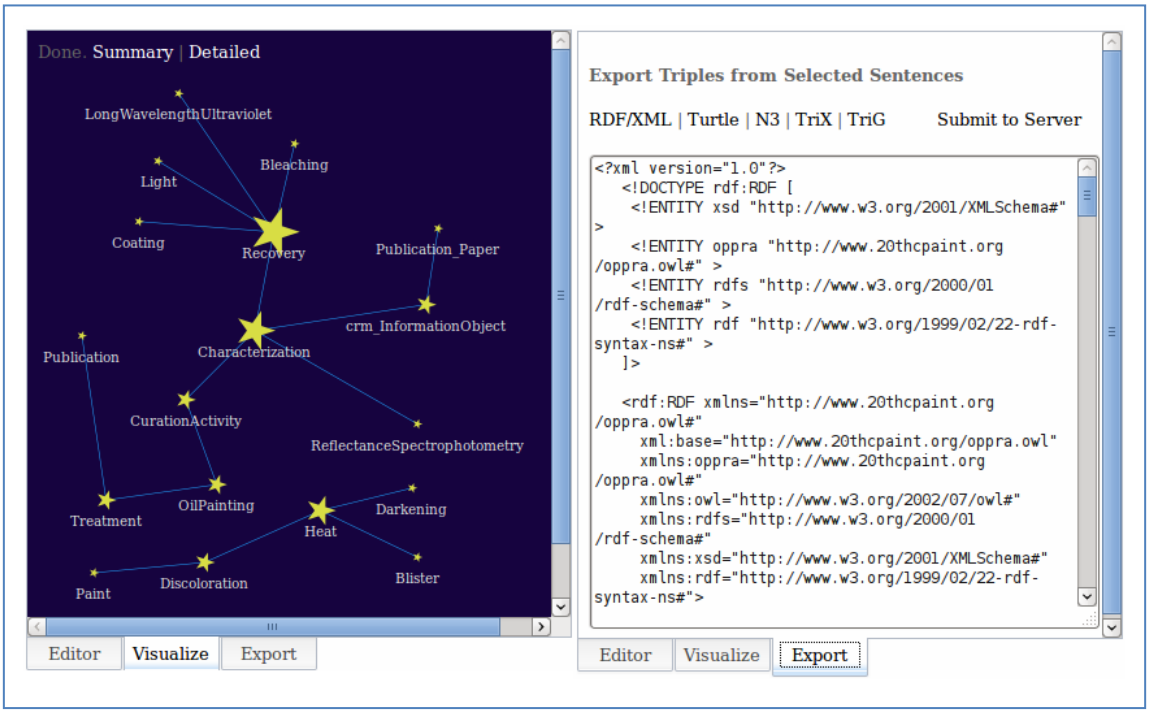



**Figure 7.6: Screenshots of the visualisation (left) and exporting/publishing (right) of RDF data**

### 7.4.3. SPARQL-based Search Interface

Figure 7.7 shows the search interface that enables SPARQL (W3C, 2008) queries across the knowledge-base. Users can search on combinations of: deterioration types, materials, and chemical compounds. For example, the query in Figure 7.7 corresponds to a search for publications that report on "*degradations of paintings that involve lead carbonate*". The results list retrieves and displays the metadata (title, authors, year of publication, publisher) for each publication that matches the query, that is, those publications about degradation (darkening, discoloration, blistering etc.) involving lead carbonate (including lead white and lead chromates).

In addition, the search results include: 1) all segments of sentences that match the query; 2) all sentences containing the matching segments; and 3) a visualisation interface that shows the matching RDF triples for each publication and the option to display the complete RDF graph corresponding to each publication. Access to the search page and some example queries are available via the Web portal (20thcpaint, 2012a).


**Figure 7.7: Search interface – results, visualisation and full sentence from original document**

## 7.5. Experimental Results

In this section, the results of performing the automatic NER and RE processing on text sentences from the corpus of publications are discussed. The performance of the NER, ambiguity resolution and RE tasks was measured by calculating: Precision (*P = Number of correctly extracted entity relations ÷ Total number of extracted entity relations*), Recall (*R = Number of correctly extracted entity relations ÷ Actual number of extracted entity relations*) and F-measure ($F = 2 \times P \times R \div (P + R)$). Calculating Precision, Recall, and F-measure is a common evaluation methodology used in text processing (Barbosa-Silva et al., 2010, Bundschus et al., 2008, Corney et al., 2004, Crowley et al., 2010, Meurs et al., 2011, Na et al., 2010, Pestian et al., 2007, Yamashita et al., 2011).

### 7.5.1. NER and Ambiguity Resolution Results

To evaluate the NER and ambiguity resolution, experiments were conducted on the 15 manually tagged publications by splitting them into 10 publications for training data and 5 publications for testing. The training and testing documents were selected randomly in a 10-fold procedure. The manual tags in the testing data were used as the benchmark for calculating *Precision (P)*, *Recall (R)* and *F-measure (F)*.

In both the training and testing procedures, the system was given a set of synonyms and their corresponding URIs from the OPPRA ontology. Altogether, 2264 synonyms associated with the OPPRA classes (as seen in Table 1) were used in the OPPRA-based gazetteer.

The number of sentence instances in the training set was 1598 sentences that included a total of 43863 tokens and 10320 lookups mapped to OPPRA URIs. In the testing phase, an evaluation process of the system includes: 1) selecting 1915 sentence instances with 45684 tokens (included in the 5 test publications), 2) applying the OPPRA-based gazetteer and NER classifier to map text segments to OPPRA URIs, and finally, 3) calculating the performance against the manually annotated terms (of 7591 lookups) in the same documents. Table 7.3 illustrates the results of the NER process.

**Table 7.3: Counts of terms/synonyms given to the training/testing data and performance of NER**

| OPPRA's Top-Level Classes | Concept Counts | | | Performance % | | |
|---|---|---|---|---|---|---|
| | **Syns** | **Manual** | **Auto.** | **P** | **R** | **F** |
| crm:Actor | **331** | **670** | **721** | **94.48** | **98.60** | **96.50** |
| Organization | 65 | 267 | 295 | 93.21 | 98.88 | 95.96 |
| crm:Person | 266 | 403 | 426 | 95.75 | 98.32 | 97.02 |
| Artist | 52 | 86 | 80 | 94.35 | 98.63 | 96.44 |
| Researcher/Conservator /Author | 214 | 317 | 346 | 97.15 | 98.01 | 97.58 |
| Source – Publication/Database | **45** | **211** | **280** | **88.15** | **91.62** | **89.85** |
| crm:Artifact | **400** | **2284** | **2197** | **92.60** | **99.06** | **95.72** |
| Painting | 16 | 49 | 35 | 91.40 | 99.75 | 95.39 |
| Device | 72 | 324 | 315 | 93.12 | 99.69 | 96.30 |
| Document – Image/Condition Report | 11 | 105 | 101 | 92.55 | 99.56 | 95.93 |
| Material – Pigment/Medium/Paint | 301 | 1806 | 1746 | 93.33 | 97.24 | 95.24 |
| Property/Attribute | **457** | **1247** | **1092** | **82.51** | **99.52** | **90.22** |
| Colour | 57 | 336 | 330 | 81.58 | 99.60 | 89.69 |
| crm:ConditionState | 193 | 326 | 321 | 82.85 | 99.41 | 90.38 |
| Measurement – Length/Energy | 207 | 585 | 441 | 83.10 | 99.55 | 90.58 |
| crm:Place | **161** | **319** | **339** | **87.16** | **88.37** | **87.76** |
| crm:Date | **189** | **503** | **631** | **94.43** | **92.23** | **93.32** |
| crm:Event | **518** | **1924** | **1791** | **91.32** | **99.43** | **95.20** |
| Treatment/Experiment/Creation | 218 | 719 | 636 | 89.95 | 99.15 | 94.33 |
| Chemical/Condition Change | 220 | 617 | 590 | 91.10 | 99.55 | 95.14 |
| Environment – Humidity/Temperature | 80 | 588 | 565 | 92.91 | 99.59 | 96.13 |
| Technique – Characterization/Artistic | **163** | **433** | **450** | **94.92** | **98.56** | **96.71** |
| **Totals** | **2264** | **7591** | **7501** | **90.70** | **95.92** | **93.16** |

### 7.5.2. RE Results

Both the rule-based and ML-based RE tasks were also evaluated using the same set of manually tagged publications for both the training and testing data. Table 7.4 shows the performance measurements for both tasks. The measurements for rule-based RE are shown based on a combination of the 4 relations (subject, hasAttribute, rdf:type and hasId) with no window and training instances. The remainder of the table shows the Precision, Recall and F-measure for 11 key relations based on their best window assigned and number of training and testing instances.

**Table 7.4: Performance measurements for rule-based and ML-based RE (the first row which does not have a window and training data), and ML-based RE (the rows with the window and training data)**

| Relation | Best Window | # Training Instances | # Testing Instance | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|
| Rule-based RE – subject, has Attribute, rdf:type, hasId | - | - | 3370 | **98.78** | **97.95** | **98.36** |
| containsMaterial | 8 | 11309 | 6914 | **78.80%** | **56.81%** | **66.02%** |
| causesCondition | 6 | 9619 | 6146 | **76.31%** | **52.54%** | **62.23%** |
| causesProcess | 11 | 12662 | 7395 | **82.12%** | **51.47%** | **63.28%** |
| employsArtifact | 10 | 12323 | 7282 | **79.90%** | **59.56%** | **68.25%** |
| carriedOutBy | 7 | 10570 | 6599 | **79.52%** | **67.20%** | **72.84%** |
| tookPlaceAt | 3 | 5010 | 3388 | **80.20%** | **74.95%** | **77.47%** |
| employsProcess | 5 | 8407 | 5491 | **71.02%** | **63.88%** | **67.26%** |
| hasTimespan | 5 | 8407 | 5491 | **75.49%** | **72.37%** | **73.90%** |
| hasAttribute | 10 | 12323 | 7282 | **85.66%** | **67.29%** | **75.37%** |
| undergoes | 15 | 13384 | 7606 | **79.76%** | **48.11%** | **60.02%** |
| actorTimePeriod | 5 | 8407 | 5491 | **82.91%** | **75.60%** | **79.09%** |

## 7.6. Analysis and Discussion

The evaluation results described in Section 7.4 revealed that the following F-measures were achieved: NER=93.16%; ambiguity resolution=90.70%; rule-based RE=98.36% and ML-based RE=60.02-79.09%. These results were an improvement on current related work in the bioinformatics field (Barbosa-Silva et al., 2010, Corney et al., 2004, Meurs et al., 2011), chemistry field (Na et al., 2010, Yamashita et al., 2011) and the biomedical field (Bundschus et al., 2008, Crowley et al., 2010, Pestian et al., 2007) using similar techniques.

Previous approaches achieved lower accuracy for specific NEs (e.g., chemicals) and relationships (e.g., chemical reactions) – focusing only on specific segments of text documents (e.g., abstracts or experimental results). For example, Corney et al. (2004) employed a template-based information extraction with a gazetteer (derived from MeSH and manually constructed thesauri) to extract relevant facts (biological information) based on a given query. Corney et. al. achieved an F-measure of 29.65% for 229 abstracts, and 47% for 130 full documents. Analysing abstracts only,

Barbosa-Silva et al. (2010) detected protein occurrences and interactions (Types 1-4) based on rules implemented for the NER and RE tasks. Their system achieved an F-measure of 60-72% for 3529 relevant abstracts, and 1957 irrelevant abstracts. The gazetteer, rule-based and ML-based methods were used by Meurs et al. (2011) to extract knowledge about *fungal enzymes*. They achieved an F-measure of 65-87%for 1493 enzymes, 984 organisms, 110 pH values, and 115 temperature values. Yamashita et al. (2011) used text mining approaches to extract information on chemical-CYP3A4 interactions from 200 abstracts. Their NER task achieved an F-measure of 89.78% and their RE task achieved an F-measure of 88.47%.

Further analysis of the results showed that the OPPRA-based gazetteer step achieved a recall of 95.92%, and the ML step achieved a recall of 100%. This indicated that the ML task (ambiguity resolution) provided better coverage than the OPPRA-based gazetteer. This was expected since the OPPRA-based knowledge-base, intentionally, did not include all of the NEs that existed in the full testing set (i.e., the 5 publications used for testing). In the future, when the knowledge-base expands by allowing users to add new instances using the *text2triples* software, the coverage of the OPPRA-based gazetteer will improve and hopefully perform as well as the ML task.

Another aspect of the NER and RE implementation is that both tasks were performed on the entire document – including titles, abstracts, figures, tables and references. This increased the errors in the results. For example, extracting the *crm:Date* in NER achieved a relatively low recall of 92.23% and a precision of 94.43%. The lower recall was because many number ranges (e.g., page numbers, figure ids, table ids) in footnotes, endnotes and figure and table captions were incorrectly tagged as a *crm:Date*.

Further analysis of the results also found that the RE task incorrectly identified/tagged relations that occurred within titles, abstracts, figure/table captions and references. Future pre-processing of each publication to identify titles, captions, tables and references and giving only related sentences to each task (classifier) would improve the accuracy of the results. For example, giving references/citations as MALLET instances for identifying the *oppra:Source* entities and relations, but not

including them within the *oppra:undergoes* classification task, would improve the RE performance.

## 7.7. Summary

In this chapter, a Web-based platform that enables the automatic extraction of structured data from textual publications about art/paint conservation is presented. The main contributions in this chapter are:

- A GATE pipeline that integrates the following tools for processing publications about paint conservation:
  - A NER tool that combines both gazetteer and ML approaches for tagging concepts and NEs within paint conservation publications;
  - A RE classifier for identifying OPPRA-based relations between NEs;
- A Web-based user interface that enables users (art conservators) to quickly and easily review, visualise and edit results graphically to ensure accurate knowledge capture;
- A SPARQL-based search interface that enables complex and detailed queries across heterogeneous full-text publications;
- A knowledge-base of facts about art/paint conservation that can easily be integrated with additional knowledge captured through further publications, databases and experiments.

Future work includes investigating methods for optimising the performance and accuracy of the automatic structured data extraction tools. For example, caching the OPPRA ontology is anticipated to improve the speed and efficiency of the OPPRA-based gazetteer. The pre-processing of publications by applying a ML model to automatically segment them (into titles, sections, figures, tables, references and footnotes) will reduce unnecessary sentence input into the NER and RE tools. Finally, as the corpus of publications (tagged with NEs and relations) expands, OPPRA will become more complete and accurate, and the OPPRA-based gazetteer is anticipated to achieve higher precision results.

Although the user interface, the triples, and the uncertainty behind making statements about the art/paint conservation were deployed within a team of 20$^{th}$

century art/paint conservators, and I.T. specialists at the University of Queensland, additional future work plans include:

- Investigating if languages other than English can be incorporated to serve a community of art/paint conservation. The functionalities of the system presented in this paper are currently available only for texts written in English. Some issues related to this aspect that could be investigated include: how much work can be estimated to extend the system functionalities to texts written in other languages? And which system components will have to be modified and/or extended?
- Evaluating the SPARQL-based search interface in order to determine if it provides better query performance and improved precision and recall over traditional publication search engines;
- Carrying out a detailed user evaluation and usability study of the system with the collaborators on the 20$^{th}$ Century in Paint project.

# Chapter 8

# SPARQL Querying and Inferencing across Local and External Databases

## 8.1. Introduction

The large volume of data being generated from the documentation of art/paint conservation activities (e.g., observations, condition assessments, characterisations, experiments, and treatments) has led to the development of numerous heterogeneous databases (both public and commercial). These databases contain information that ranges from artists' biographies and their techniques, to information on paint materials and chemistry, degradation mechanisms, characterisation techniques, experimental results, and cleaning/conservation methods. Relevant databases also contain a wide variety of data types including textual reports, images and file formats associated with the characterisation or analysis of paint materials (e.g., spectra, electron backscatter images, X-ray images, and near infrared light images).

In addition, as described in Chapter 7, significant prior research in art/paint conservation has been published in traditional publications. The knowledge in these publications is difficult to discover and retrieve because it is distributed across repositories and publishing houses, embedded and hidden within large amounts of unstructured text, and expressed using a wide variety of different terminologies.

As a result, today's art conservators and materials scientists are confronted with significant material-based preservation questions, but lack the integrated knowledge-base to inform their decision-making. As described in Sections 1.1.1, 1.4, 1.6, and 3.6, conservators and materials scientists are demanding data management and integration tools that enable them to search across these disparate databases, and to correlate their own organisation's characterisation and experimental data sets with external, publicly available data, in order to identify the causes of art/paint preservation issues and determine the optimum treatments.

The current distributed, unstructured, and heterogeneous nature of relevant data makes it extremely difficult for conservators to search and aggregate information to find answers to the problems that they face. For example, consider the question "*What additives cause paint instability?*" To answer this, the conservator needs to search paint databases (e.g., W&N), find what additives are used (e.g., aluminium stearate), and then search chemical databases (e.g., CAMEO) for each additive's physical and chemical properties. The objective is to determine the effects of specific additives on paint materials (e.g., chemical reactions between pigments, oils and additives) and the effect of other environmental parameters (e.g., humidity, temperature, UV light) on such chemical reactions. In addition, as demonstrated in Chapter 7, relevant publication archives (e.g., JAIC, JSTOR Studies in Conservation, Analytical Chemistry and AICCM Bulletin) also provide valuable information about these paints and additives, but the task of searching and retrieving information from the publications within these archives is extremely tedious. Chapter 7 has illustrated how the relevant data can be extracted and stored as RDF in a standardised format. This chapter illustrates how the extracted RDF structured data can now be exploited by enabling its integration with other databases through a SPARQL query interface.

This chapter describes the Data Aggregation and Linking Interface– DALI for 20$^{th}$ century art/paint conservation information. DALI aims to address the data integration requirements identified from the workshops held for the APTCCARN members (Sections 1.1.1, 1.4, 1.6, and 3.6). More specifically, DALI aims to enable conservators and scientists to specify queries (such as those that were identified in Section 3.5) and retrieve responses to these queries.

Specifically, the aim of DALI is to provide a federated search interface over key information sources for art/paint preservation that have been integrated through the underlying OPPRA ontology (described in Chapter 4). The integrated datasets include: experimental databases from the 20th Century in Paint project (described in Chapter 6 – Sidney Nolan Paint Archive and Mecklenburg Samples); structured data extracted from past publications using *text2triples* (described in Chapter 7); and records from the following relevant publicly available databases (that were identified as useful by art conservation experts in the 20th Century Paint project):

- W&N (2009) – detailed information on the manufacture of pigments, binders, mediums and paints used by nineteenth century painters;
- DAAO (2010) – biographical data about Australian artists, designers, craftspeople and curators;
- IRUG Spectral Database (IRUG, 2010) – a forum for the exchange of infrared and Raman spectroscopic information, reference spectra and materials;
- CAMEO (MFA-Boston, 1997) – a searchable information resource developed by the Museum of Fine Arts, Boston, containing chemical, physical, visual, and analytical information on historic and contemporary materials used in the production and conservation of artistic, architectural, archaeological, and anthropological materials;
- Forbes Pigment Database (MFA-Boston, 2010) – a collection of colorants (assembled by Edward Waldo Forbes) that have been analysed widely. This collection aims to provide one central, searchable and readily-accessible compilation of information on pigments;
- Color of Art Pigment Database (Myers, 2010) – an artists' paint and pigments resource with colour index names, pigment codes, colour index numbers and chemical composition;
- FT-IR Spectra of Binders and Colorants (Vahur, 2009) – a selection of infrared spectra of various paint and coating materials registered at the University of Texas Testing Centre and Department of Chemistry;
- NIST Chemistry WebBook (NIST, 2011) – data (compiled and distributed by NIST under the Standard Reference Data Program) that contains thermo-chemical, reaction thermo-chemistry, IR/Mass/UV spectra, gas chromatography, constants of diatomic molecules, ion energetic, and thermo-physical properties;

- Paint and Ink Formulations Database (Flick, 2005) – provides the seminal paint and ink formulations compiled by Ernest Flick that were published during the last decade.

These datasets have been chosen because of their ready availability and range of relevant content. However, DALI has been designed so that additional datasets can easily be incorporated in the future (as required by the 20th Century in Paint project teams or as they become available) by applying one or more of the following methods to incorporate the new database: SPARQL, REST APIs, Web crawling, or database/RDF mapping (e.g., D2R (Bizer and Cyganiak, 2006)).

In addition to data integration, DALI applies reasoning over the aggregated data and infers new facts (i.e., implicit knowledge). Reasoning and inferencing are implemented using a set of OWL 2 RL rules (described in Chapter 4). Details of the inferencing implementation are described in Section 8.3.4.

## 8.2. Related Work

This section covers two topics relevant to the application of Semantic Web techniques to data management for art conservation. These are: ontology-based data integration; and ontology-based reasoning and querying.

### 8.2.1. Ontology-based Data Integration

Below are described significant related research efforts that have leveraged Semantic Web technologies to integrate data across museums and art galleries (Aliaga et al., 2011, Hyvönen et al., 2009, Binding, 2010, Binding et al., 2008, Hyvönen et al., 2006, Mellon, 2009, Monroy et al., 2010, Toledo et al., 2009, Tudhope et al., 2011, Vlachidis, 2012, Vlachidis et al., 2013). The examples discussed in this section represent the most significant or innovative projects: ConservationSpace, CultureSampo, a Brazilian indigenous cultural heritage proposal, Semantic Technologies for Archaeological Resources – STAR, Semantic Technologies Enhancing Links and Linked data for Archaeological Resources– STELLAR, English Heritage Centre for Archaeology data integration project, the AMA project, and the DECHO project.

*ConservationSpace*

ConservationSpace (Mellon, 2009) is a Mellon-funded project that aims to convert the British Museum collection's metadata into RDF that complies with CIDOC-CRM. The aim is also to support metadata extensibility so that data/metadata from other museums (and similar funded projects such as those mentioned in Section 2.3) can gradually be incorporated. The scope of ConservationSpace includes an RDF gateway that allows the following deliverables: a) import and export services to reduce overheads and allow institutional control of online and offline data; b) search and access mechanisms including inference and relation navigation; c) standard Web-based creation, modification and deletion of institutional data; d) CIDOC-CRM-controlled user interface (and an option for uncontrolled comments) for image annotation with different zooming levels; e) image comparison service through different layouts, transparency and pixel comparison; f) relation/link editor through controlled vocabularies and navigation services; and g) visualisation services for spatial and temporal data. ConservationSpace uses CIDOC-CRM as the common model for data integration and does not focus specifically on painting conservation.

Following the ConservationSpace efforts, the conceptual model, and the technical framework were implemented within the ResearchSpace project (Alexiev et al., Oldman, 2010, Oldman et al., 2014). The implementation uses OWLIM for the CIDOC-CRM-based triple store. OWL 2 RL is also used to provide inferencing, and enhance the data search and population. A total of 120 rules were implemented. These are: rules that implement RDFS reasoning within the default OWLIM (14 rules); and rules that implement methods for conjunctive (e.g., checking the type of a node), disjunctive (parallel), serial (property path), and transitive reasoning (106 rules). The OWLIM triple store includes RDF mapped from the following datasets: the Europeana Data Model (EDM) repository; CLAROS (Kurtz et al., 2009, OeRC, 2014); and the Poznan Supercomputing and Networking Center (PCSS, 2014).

*CultureSampo*

CultureSampo (Hyvönen et al., 2009, Hyvönen et al., 2006) is a platform that aimed to combine and access heterogeneous archives of cultural heritage-related content within the MuseumFinland Web portal. Each metadata schema used to represent data was mapped onto a shared ontology (the ONKI ontology). CultureSampo

generalises MuseumFinland in the following ways: a) cross-domain heterogeneous content of virtually any form (e.g., images, narrative stories and historical events); b) event-based knowledge representation for the implicit knowledge embedded in the integrated content; and c) collaborative content creation using Web 2.0 techniques.

*Brazilian Indigenous Cultural Heritage Proposal*

A Semantic Web approach for sharing resources between different Brazilian indigenous cultural heritage institutions was proposed by Toledo et al. (Toledo et al., 2009). The specific goals for this proposal include: a) integrated data from different museums and institutions for indigenous cultural heritage in Brazil; b) an extensible ontology for Brazilian indigenous cultural heritage; and c) building knowledge about Brazilian indigenous cultural heritage using Wikis.

*STAR*

STAR (Binding, 2010) aimed to address the issues concerning the extraction and representation of time period information by exploiting the potential of a standard ontology for cultural heritage. It extended an ontology designed for the archaeology excavation and analysis process. This ontology was then used to link digital archive databases, vocabularies and associated literature. Temporal events in the ontology were defined to include: *intervalEqual*, *intervalBefore* and *intervalAfter*.

*STELLAR*

The STELLAR project (Tudhope et al., 2011, Vlachidis, 2012, Vlachidis et al., 2013) addressed the problematic issue of mapping terms from the CIDOC-CRM (and its extension) to time periods (temporal events). STELLAR provided more support and guidance to data providers and generalized the data mapping/extraction techniques to help third party data providers undertake this work. STELLAR aimed at making the mapping/extraction process easier for data providers (who are familiar with their own data, but less familiar with the ontology).

*English Heritage Centre for Archaeology Data Integration*

A data integration approach using the English Heritage Centre for Archaeology ontological model called CRM-EH (an extension to the CIDOC-CRM) was proposed by Binding et al. (2008). The aim of this project was to demonstrate the potential

benefits of integrating and searching across institutional data expressed as RDF and that conformed to a common overarching conceptual data structure schema.

*The AMA Project*

The interoperability of cultural heritage datasets and schemas between different platforms available on the Web was exploited by the AMA project. The AMA project is part of the EPOCH project (Eide et al., 2008). The tools developed within the AMA project are aimed at providing the semi-automated mapping and integration of cultural heritage custom data to the CIDOC-CRM. It also aims to provide a semantic framework to store, manage and browse the encoded information via user-friendly interfaces (Eide et al., 2008, Hernández et al., 2008, Monroy et al., 2010).

*The DECHO Project*

DECHO (Aliaga et al., 2011) is a Semantic Web framework designed to support the acquisition, management and visualisation of archaeological data. The data acquisition component aims to support the fast, easy and accurate addition of 3D object models and factual data, including narrations (disseminating knowledge throughout communities of different users from students to experts). Using their ontology management system, a two-layer abstraction (conceptual mapping, with machine level mapping – e.g., cidoc-crm:E55.Type with '*craftsman's signature*') enables fast and intuitive access to a heterogeneous set of data sources.

The majority of related projects above use Semantic Web technologies and principally the CIDOC-CRM ontology to assist with the management and understanding of cultural heritage artefacts. Past research primarily focuses on the physical provenance of artworks (e.g., art history) and the linking of physical provenance and historical contextual information to digital representations of cultural artefacts. Moreover, semantic inferencing primarily involves sub-classing rules that associate upper level classes (e.g., Entity, Artifact, Agent, Event, Attribute, Date, Place, Material, and Document) in each given domain. The research described in this thesis is unique because it is the first that focuses on the application of Semantic Web technologies to the conservation of artworks (and more specifically twentieth century paintings). It is the first that aims to link information about art history and

artistic techniques with information about paint chemistry, paint analysis, experimental data and past publications.

### 8.2.2. Ontology-based Reasoning and Querying

This section provides an overview of significant related research efforts that have leveraged Semantic Web technologies to reason over, and query data across museums and art galleries (Aroyo et al., 2007, Aliaga et al., 2011, Hyvönen et al., 2009, Barak et al., 2009, Koutsomitropoulos and Papatheodorou, 2007, Kurtz et al., 2009, Hyvönen et al., 2006, Haslhofer et al., 2010, OeRC, 2014, Monroy et al., 2010, van Ossenbruggen et al., 2007, Sanderson and Van de Sompel, 2010, Schmidt et al., 2011, Theodoridou et al., 2010, Toledo et al., 2009, Wielemaker et al., 2008). The examples discussed in this section are: MOSAICA, MultimediaN E-Culture, ClioPatria, Expressive Reasoning about Cultural Heritage Knowledge Using Web Ontologies, Modelling and querying provenance by extending CIDOC CRM, Europeana, CLAROS, and Recovering Brazilian Indigenous Cultural Heritage Using New Information and Communication Technologies.

*MOSAICA*

MOSAICA (Barak et al., 2009) provided a generic framework for users to actively engage in preserving their heritage via activities such as investigation, exploration and storytelling. MOSAICA aimed to develop a toolbox of generic technologies for the preservation of cultural heritage resources (e.g., photos, documents, video and sound). The current state of the system includes: a) an ontology-based search tool across the integrated sources (e.g., Events, Notions, People, Periods, Places, Resources, and Things); b) the ability for users to select a geographic location and display it in a map using an ontology taken from BUSTER (a system developed at the University of Bremen to integrate and query heterogeneous information from different geospatial datasets); and c) virtual objects in the form of Web-based resources (e.g., a story-telling system).

*MultimediaN E-Culture*

MultimediaN E-Culture (Aroyo et al., 2007, van Ossenbruggen et al., 2007) brought together multiple online cultural heritage repositories in the Netherlands in a way that is comparable to the MuseumFinland project (Hyvönen et al., 2009, Hyvönen et al., 2006, Hyvönen et al., 2005). It was aimed at public users and non-technical researchers with a generic browser (the CHIP browser) to explore the databases through any facet (e.g., artist, genre or period). The CHIP browser (which drew on their defined ontological mapping between the individual datasets and the Getty thesauri) resulted in: a) providing automated artwork suggestions via the '*ArtworkRecommender*' based on users' ratings; and b) personalised tours of the Rijksmuseum's artworks (that can be downloaded to a handheld device).

*ClioPatria*

ClioPatria (Wielemaker et al., 2008) is a Prolog framework for constructing Semantic Web applications. It is an open-source system that provides APIs for scalable semantic graph searches (a re-usable core of the E-Culture demonstrator), with backward chaining inferencing. It integrates the SWI-Prolog libraries for RDF and HTTP services into a Semantic Search Web Server.

*Expressive Reasoning about Cultural Heritage Knowledge Using Web Ontologies*

Koutsomitropoulos and Papatheodorou (2007) take CIDOC-CRM (Version 3.4) and convert it to an OWL representation in order to extract knowledge using inferencing. Their approach defined the semantics between time periods such as '*overlaps*', '*precedes*' and '*follows*' and inferred relationships via temporal reasoning.

*Modeling and querying provenance by extending CIDOC CRM*

Another extension to the CIDOC-CRM ontology to capture the modelling and query requirements regarding the provenance of physical and digital objects was proposed by Theodoridou*et et al.* (2010). In this extension, a number of indicative provenance query templates for various domains were developed using Semantic Web technologies.

*Europeana*

The Europeana tools (Haslhofer et al., 2010, Sanderson and Van de Sompel, 2010, Schmidt et al., 2011) build on previous projects for European cultural heritage content integration: ARTISTE, eCHASE, SCULPTEUR and mSpace (Addis et al., 2006, Goodall et al., 2004, Smith et al., 2005) to provide cross-archival search capabilities for galleries using RDF metadata. Europeana uses an ontology-driven approach (CIDOC-CRM with several extensions) to provide adaptive search and visualisation mechanisms for 2D and 3D objects. Data types in this project include digital images, 3D models, associated metadata, free-text documents and numerical tables.

*CLAROS*

Similar to the Europeana project, CLAROS (Kurtz et al., 2009, OeRC, 2014) is an international research collaboration, using the latest information and communication technologies (ICT) to enable simultaneous searching of major collections in universities, research institutes and museums. It is a multi-domain project that provides ways to link geographically distributed artefacts (e.g., paintings, drawings, sculptures, coins, eastern ceramic, western ceramic, aerial photographs, and inscription) via semantic web technologies. The metadata from each content provider is mapped to CIDOC-CRM and stored in a common RDF triple store with a SPARQL-search interface.

*Recovering Brazilian Indigenous Cultural Heritage Using New Information and Communication Technologies*

The Semantic Web approach proposed by Toledo *et al*. (2009) provides the Brazilian cultural heritage community with the following services (in addition to the data integration tools): a) tools that enable archaeologists to visualise artefacts in 3D, identify and catalogue artefacts, virtually re-construct broken ceramics into whole objects, and enter information about artefacts; b) customised tours according to specific themes and user profiles; and c) user-selective participation in exhibits through electronic books and social networks.

Furthermore, most of the projects reviewed in Sections 6.2, 7.2, and 8.2.1 (that have used Semantic Web technologies) have aimed to provide querying and visualisation capabilities for the cultural heritage community. A review of the related literature above reveals that there has been no previous research that focuses on the application of ontology-based data integration and semantic searching, querying and inferencing to provide new knowledge and answer complex queries specifically related to art/paint conservation.

## 8.3. Data Aggregation and Linking Interface – DALI

DALI aims to enable researchers to access relevant disparate art/paint preservation knowledge (e.g., paint chemistry data, deterioration mechanisms and characterisation/imaging data) via a single Web-based search interface. The implementation methodology is based on the following steps:

- Populating the OPPRA RDF triple store with instances from:
  - Internal experimental data – Sidney Nolan Paint Archive, and Mecklenburg Samples;
  - Structured information from past publications – *text2triples*;
  - Public databases – W&N, DAAO, IRUG Spectral Database, CAMEO, Forbes Pigment Database, Color of Art Pigment Database, FT-IR Spectra of Binders and Colorants, NIST Chemistry WebBook, and Paint and Ink Formulations Database;
  - Semantic inferencing – OWL 2 RL profile.
- SPARQL (W3C, 2008) querying over the OPPRA-based RDF triple store;
- Implementing a user interface that supports simple keyword, advanced (Boolean) search and SPARQL queries and returns results (e.g., customised result and graph visualisations) with links to data sources (e.g., record, sentence) that match the query.

### 8.3.1. Data Model

The data aggregation and inferencing described here is based on the OPPRA ontology (described in detail in Chapter 4). The OPPRA ontology enables cross-disciplinary queries to be performed against the information extracted from local and external datasets (e.g., experimental data, *text2triples* knowledge, paintings, artists,

artistic techniques, materials, manufacturers, condition states, degradation mechanisms, characterisation data, and conservation treatments).

Figure 8.1 illustrates an example of a record (id: 43) modelled on the OPPRA ontology. The record provides information on a sample (#43) that is taken from DULUX black paint (brand: *BALM (Australia) Pty. Ltd*; code: *44146*; and binder: *alkyd*). The OAI-ORE representation of *SidneyNolanArchive* and *Record43* shows that the Sidney Nolan Paint Archive *aggregates* record 43, and that record 43 *aggregates* 5 statements (*oppra:Statement*). Each statement is represented as a Named Graph (*context* in the OpenRDF Sesame Triple Store), and it includes the triple (subject, predicate, object) associated with the given knowledge (e.g., Sample43 *undergoes* Py-Gc-Ms, and Py-Gc-Ms *outputs* SST.tiff). Py-Gc-Ms is short for Pyrolysis-gas chromatography-mass spectrometry. This is a characterisation method that generates a TIFF image. PXRF is Portable X-Ray Fluorescence.

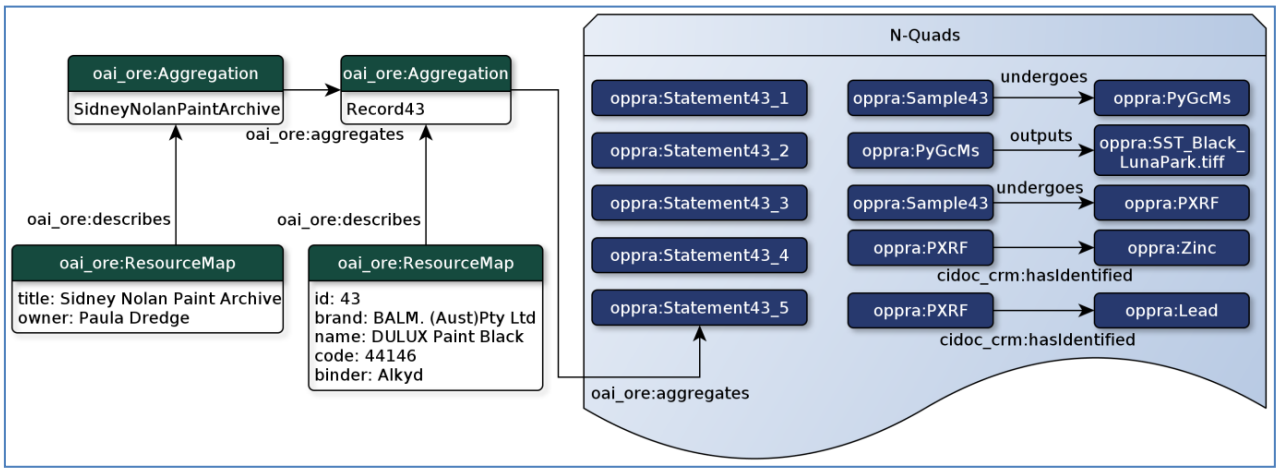



**Figure 8.1: RDF graph of characterisation metadata on a Sidney Nolan Paint Archive sample "43" which has undergone Pyrolysis-gas chromatography-mass spectrometry to generate a TIFF image. It has also undergone Portable X-Ray Fluorescence that indicates the presence of both Zinc and Lead.**

### 8.3.2. Data Integration

Data extracted from multiple sources (local databases, publications, external datasets, and semantic inferencing) using different Java scripts and classes, is stored in the OPPRA RDF triple store. This section provides details (with examples) of the following steps: RDF conversion from local databases, RDF storage of past publications, and RDF conversion and extraction from external datasets (in

particular, Web crawling of the W&N archive, and SPARQL querying the NIST Chemistry WebBook).

*RDF Conversion from Local Databases*

When users upload experimental data (sample information, experimental details, experimental results, characterisation data, etc.) to the 20th Century in Paint project database as a new project (e.g., the Sidney Nolan Paint Archive project and the Mecklenburg Samples project), the data is stored in a content management system – Jackrabbit (Apache, 2004). Jackrabbit stores the uploaded data as binary files, and associates these files with metadata that describes their content. Metadata is then extracted and converted to valid OPPRA triples and stored in the OWLIM (Bishop et al., 2012) OpenRDF Sesame triple store.

For example, the RDF graph shown in Figure 8.2 records the information associated with the results of applying μ-Raman characterisation to a sample "001" by conservator "Gillian Osmond". An example of the actual Trig export (Named Graphs/contexts, and N3 triples stored in the OPPRA-based RDF triple store) is provided in Chapter 6 (Section 6.3.1).

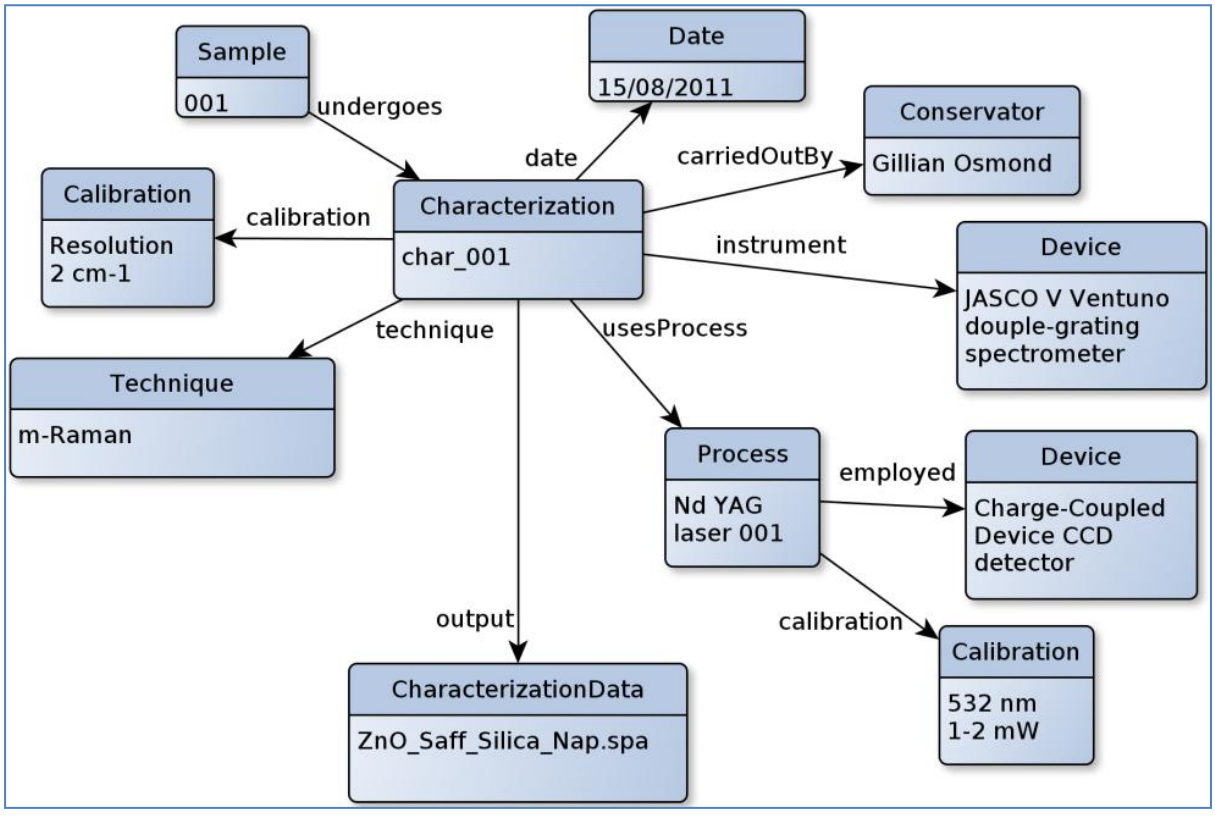



**Figure 8.2: RDF graph of characterisation metadata on a Mecklenburg Samples record "001"**

*RDF Storage of Past Publications*

As described in Chapter 7, data is extracted from past publications and stored in the OPPRA RDF triple store via the following steps:

- A set of publications about art/paint conservation (described in Chapter 7) is acquired to provide a corpus of relevant knowledge;
- Structured knowledge (RDF triples) is extracted from the relevant publications using the *text2triples* software (described in Chapter 7). The software combines GATE (Cunningham et al., 2011) and MALLET CRF sequence tagging (McCallum, 2002) to generate a semi-automated framework for the structured data extraction;
- The structured knowledge generated by *text2triples*is verified, corrected where necessary, and stored in the OPPRA RDF triple store (OWLIM implementation).

Section 7.4.2 (Chapter 7) provided an example of one publication included in the corpus, and the resulting set of RDF statements extracted from this publication:

- *Monico, L, Van der Snickt, G, Janssens, K, De Nolf, W, Miliani, C, Dik, J, Radepont, M, Hendriks, E, Geldof, M &Cotte, M 2011, 'Degradation Process of Lead Chromate in Paintings by Vincent van Gogh Studied by Means of Synchrotron X-ray Spectromicroscopy and Related Methods. 2. Original Paint Layer Samples', Analytical Chemistry, vol. 83, no. 4, pp. 1224-31* (Monico et al., 2011).

*RDF Conversion and Extraction from External Datasets*

This step involves the extraction of knowledge from external relevant datasets (identified by art conservation experts in the 20$^{th}$ Century in Paint team) and its conversion to RDF graphs, that comply with the OPPRA ontology. For each dataset, an optimum entry point (e.g., SPARQL query, REST API or keyword search) is identified to retrieve records for further processing/RDF conversion.

Currently, selected records from W&N (2009), DAAO (2010), IRUG Spectral Database (IRUG, 2010), CAMEO (MFA-Boston, 1997), Forbes Pigment Database (MFA-Boston, 2010), Color of Art Pigment Database (Myers, 2010), FT-IR Spectra of Binders and Colorants (Vahur, 2009), NIST Chemistry WebBook (NIST, 2011) and

Paint and Ink Formulations Database (Flick, 2005) are incorporated. These datasets have been chosen because of their ready availability and range of relevant content.

However, additional datasets can easily be incorporated – by identifying the entry point to the database that will be incorporated (e.g., SPARQL API), retrieving the records (or fields) that are of interest, converting them to RDF (e.g., D2R (Bizer and Cyganiak, 2006)) and storing them in the OPPRA RDF triple store.

The following discussion provides two examples of the data extraction step: Web crawling of the W&N archive; and SPARQL querying of the NIST Chemistry WebBook.

The W&N Archive provides digital recipes of pigment, paint, varnish and oil from the 19th century archive of Winsor and Newton. Data is extracted from this database by searching for classes and their synonyms that are defined in the OPPRA ontology (e.g., *zinc, varnish, linseed, and watercolour*) using the "*Search Index*" URL in the W&N public portal. The search results are provided as an HTML table (that specifies recipe names – original and interpretation, topics, materials and years). This data is processed and mapped to RDF triples compliant with OPPRA. Relations are assigned based on the data in each row. The Named Graph below for example, indicates that a record "*WNRecord_DRP001AL01*" with the given W&N URL contains data on a "*Sample*" of "*Drying Linseed Oil*", and the sample underwent a specific "*Activity*" that took place in "*1980*".

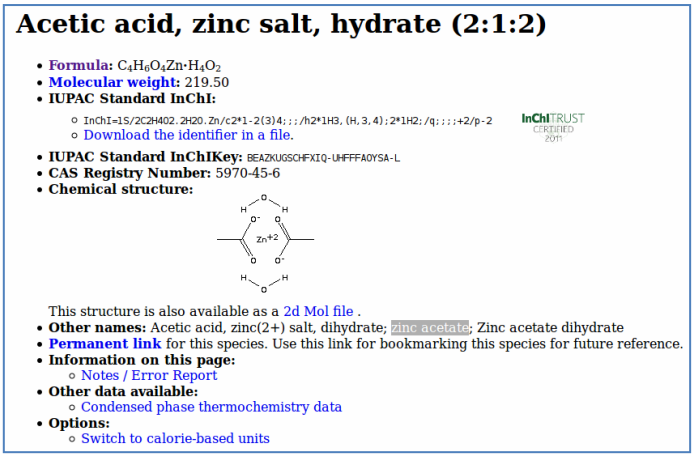
```
@prefix : <http://www.20thcpaint.org/oppra.owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
:WinsorAndNewtonArchive {
:WinsorAndNewtonArchiveaowl:NamedIndividual , :Database ;
:aggregates :WNRecord_DRP001AL01 .
}
:WNRecord_DRP001AL01 {
:WNRecord_DRP001AL01 aowl:NamedIndividual , :Record ;
:hasURL "www.hki.fitzmuseum.com/arch.php?u=DRP001AL01" .
:DryingLinseedOilaowl:NamedIndividual , :LinseedOil ;
rdfs:label "Drying Linseed Oil" ; :undergoes _:Activity1980 .
 _:Activity1980 :hasTimespan :Timespan_1980 .
:Sample_DRP001AL01 aowl:NamedIndividual , :Sample ;
```

```
rdfs:label "Sample DRP001AL01" ; :hasId "DRP001AL01" ; :takenFrom
:DryingLinseedOil .
}
```

The NIST Chemistry WebBook provides access to data compiled and distributed by NIST under the Standard Reference Data Program (NIST, 2012). The online SPARQL API is used to retrieve a set of records (based on SPARQL queries constructed specifically for various case studies in the 20th Century in Paint project), and store the records in the OPPRA RDF triple store. For example, Figure 8.3 shows a construct-based SPARQL query that gathers information on "*Zinc Acetate*". The screenshot on the right shows the NIST-based results of the query. The Named Graph below represents the statements for the record URL, chemical structure, and synonyms of zinc acetate "$C_4H_6O_4Zn \cdot H_4O_2$" extracted by performing this query.

```
PREFIX rdfs:
<http://www.w3.org/2000/01/rdf-schema#>

Construct {?s ?p ?o} where {
  ?s ?p ?o .
  ?o rdfs:label ?label .
  filter regex(
     ?label,
"zinc acetate",
     "i"
  )
}
```



**Acetic acid, zinc salt, hydrate (2:1:2)**

- **Formula:** $C_4H_6O_4Zn \cdot H_4O_2$
- **Molecular weight:** 219.50
- **IUPAC Standard InChI:**
  - InChI=1S/2C2H4O2.2H2O.Zn/c2*1-2(3)4;;;/h2*1H3,(H,3,4);2*1H2;/q;;;;+2/p-2
  - Download the identifier in a file.
- **IUPAC Standard InChIKey:** BEAZKUGSCHFXIQ-UHFFFAOYSA-L
- **CAS Registry Number:** 5970-45-6
- **Chemical structure:**

This structure is also available as a 2d Mol file .

- **Other names:** Acetic acid, zinc(2+) salt, dihydrate; zinc acetate; Zinc acetate dihydrate
- **Permanent link** for this species. Use this link for bookmarking this species for future reference.
- **Information on this page:**
  - Notes / Error Report
- **Other data available:**
  - Condensed phase thermochemistry data
- **Options:**
  - Switch to calorie-based units

```
@prefix : <http://www.20thcpaint.org/oppra.owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
:NISTChemistryWebBook {
:NISTChemistryWebBookaowl:NamedIndividual , :Database ;
:aggregates :NISTRecord_AceticAcidZincSaltHydrate .
}
:NISTRecord_AceticAcidZincSaltHydrate {
:NISTRecord_AceticAcidZincSaltHydrateaowl:NamedIndividual , :Record ;
:hasURL"http://www-
http://webbook.nist.gov/cgi/cbook.cgi?ID=C5970456&Units=SI" .
:C4H6O4Zn.H4O2 a owl:NamedIndividual , :ChemicalStructure ;
rdfs:label "C4H6O4Zn.H4O2" ; :hasSynonym "acetic acid, zinc(2+) salt,
dihydrate", "zinc acetate", "zinc acetate dihydrate" .
}
```

**Figure 8.3: SPARQL construct query that converts NIST Chemistry WebBook result to an OPPRA graph**

### 8.3.3. Entity Resolution

A common problem that arises when integrating disparate legacy databases is entity resolution. This section describes the entity resolution process that is used to link databases via common entities that are uniquely identified by URIs. The process determines when a particular entity in one database is the same as a particular entity in another database or publication, and links the two entities and their associated related data/metadata via a single URI.

To resolve entities in DALI, manual intervention is firstly needed to define the fields/attributes associated with each OPPRA class which need to be compared to determine if two entities are the same. The defined fields are the containers (e.g., tables' columns, SPARQL results' variables) that have the text to be resolved (e.g., labels, synonyms). Suppose there is a table that has a list of artist names, artist aliases, and artworks painted by these artists. To extract the OPPRA statements *paints(Artist, Painting)*, the following procedure is performed to extract such statements for each row in the table:

- For each row (artist) in the table, create a list of *synonyms*, and a list of *artworks*;
- Initiate the variables *artist*, and *painting*;
- For all *synonyms*, find a possible URI (in OPPRA) by executing the SPARQL query: *select distinct ?uri where{?urioppra:hasSynonym "synonym"}*;
- If a *uri* is found, then *artist=uri*, or else create a *uniqueName* from the first synonym, and *artist="oppra:"+uniqueName*;
- For each *artwork*, find a possible URI (in OPPRA) by executing the SPARQL query: *select distinct ?uri where{?urioppra:hasSynonym "artwork"}*;
- If a*uri* is found, then *painting=uri*, or else create a *uniqueName* from *artwork*, and *painting="oppra:"+uniqueName*. The statement *paints(artist, painting)* can then be added to the knowledge-base.

DALI is flexible enough to find name variants using the *oppra:hasSynonym* property, but sufficiently restrictive to produce a manageable candidate list despite being a large-scale knowledge-base. However, the provision of automatic entity resolution remains a significantly challenging task.

Despite good progress in entity resolution that leads to levels of performance close to manual results, such methods have not performed so well in the art/paint

preservation domain (Chieu and Teow, 2012, Kim et al., 2004, Krallinger et al., 2013, Liao and Zhang, 2012, Zhang and Elhadad, 2013). Challenges arise, for example, due to variations in: how an entity may be referenced (e.g., '*Sydney Nolan'* and '*the artist' (Nolan)*); from the existence of several entities with the same name (e.g., '*Sydney Nolan'* and '*S. Nolan'*); or even from spelling mistakes in the name. Name disambiguation remains a major challenge in the cultural heritage domain, and more specifically in the art domain.

Further manual processing in the newly created URIs can also be done to expedite the linking task to existing URIs (if applicable), by using the OWL 2 functional property *owl:sameAs*. Using this property (*sameAs(a,b)*), an inferencing engine assigns all relations (annotations, metadata, object properties, and data properties) that belong to an instance *a* to the instance *b*, and vice versa. An automatic solution that can detect similarities between different URIs in the knowledge-base has not been investigated in this thesis; however, it is an open challenge that is worth investigating in the future.

### 8.3.4. Semantic Inferencing across the Knowledge-base

As demonstrated in Chapter 4, inferencing is applicable to a number of aspects of art/paint preservation, including the relationships between:

- Causes of degradation and condition states of materials;
- Characterisation techniques, instrument use, and data outputs;
- The creation of materials/artefacts, corresponding actors, artistic techniques, periods, and locations (temporal relations);
- The physical and digital provenance of artefacts, and the temporal/spatial representation of data (e.g., timelines and maps).

In the work described here, inferencing rules are applied over the OPPRA RDF triple store using the OWL 2 Rule Language (Motik et al., 2012) profile. These rules are executed using OWLIM (Bishop et al., 2012) – which extends the OpenRDF Sesame triple store by adding the OWL 2 RL inferencing profile.

Some specific examples of inferencing rules, and instances of extracted knowledge include:

- **Transitive properties**: Sample001 *consistsOf* Pb_2_O_3; Pb_2_O_3 *consistsOf*Pb → Sample001 *consistsOf*Pb;
- **Temporal relations, paintedOn→wasPaintedBy .hasTimespan**: Painting_The_Journey was*paintedBy* PaintingActivity001; PaintingActivity001 *hasTimespan* Timespan_1943-1992 → Painting_The_Journey *hasTimespan*Timespan_1943-1992;
- **Contains relations, indicates ← indicated .consistsOf**: SEM_Activity *indicates* Zn_O. Zn_O *consistsOf* Zn → SEM_Activity *indicates* Zn;
- **Semantic relations, refersTo ← outputBy . indicates**: Spectra_IR8897 *outputBy*IR_Spectroscopy . IR_Spectroscopy*indicates* Alizarin → Spectra_IR8897 *refersTo* Alizarin.

## 8.4. System Architecture

Figure 8.4 shows the high-level architecture of the DALI system which comprises a set of key components on both the server and client sides. The design of DALI was based on a decision to adopt Web 2.0 technologies (AJAX and Web services) to enable the fast and flexible development of a user-centric application that provides real-time access to dynamically changing datasets.

The user interface sits on the client side and:

- Is rendered by the Dojo Toolkit (Dojo-Foundation) that is designed to enable rapid development of AJAX-based applications and Websites;
- Uses the SPARQL JavaScript Library that supports querying of the OPPRA ontology on the server side (Feigenbaum et al., 2006).

The following key components run on the server side:

- The knowledge-base which consists of the OPPRA ontology, and the RDF instances stored in OWLIM (Bishop et al., 2012) which uses the OpenRDF Sesame triple store (Aduna, 1997) and the OWL 2 RL inferencing profile (Motik et al., 2012);
- Jackrabbit repositories for the 20$^{th}$ Century in Paint project datasets (experimental data associated with sub-projects) – metadata in the repository is transformed into RDF which is stored in the OPPRA-based RDF triple store;

- The *text2triples* framework that extracts structured data from past publications, and saves, exports and visualises this structured knowledge (*statements*) to/from the OPPRA RDF triple store;
- Various other scripts implemented in Java – these scripts perform data extraction from public databases and transform the extracted data into RDF triples stored in the OPPRA RDF triple store;
- Inferencing rules that are implemented using OWL 2 RL (Motik et al., 2012) and applied to extract new facts (hidden art/paint preservation knowledge) from the integrated data sets (the complete set of OPPRA RDF triples).

## 8.5. SPARQL Queries

Searching across the given datasets involves the use of the SPARQL Query Language for RDF (W3C, 2008) – a W3C recommendation that is able to retrieve and manipulate data stored in RDF format. In addition, OWLIM supports querying of the knowledge-base using SPARQL via the REST API which enables searching of the data with client-side libraries such as AJAX and SPARQL.

For example, the following SPARQL statement represents the query "*show experiments about cleaning blanched artworks with mineral spirits*":

```
PREFIX oppra:<http://www.20thcpaint.org/oppra.owl#>

select distinct ?experiment where{?artwork oppra:undergoes ?experiment .
?artwork oppra:hasConditionState oppra:Blanched . ?experiment
hasDescription oppra:Cleaning; oppra:usesMaterial oppra:MineralSpirit}
```

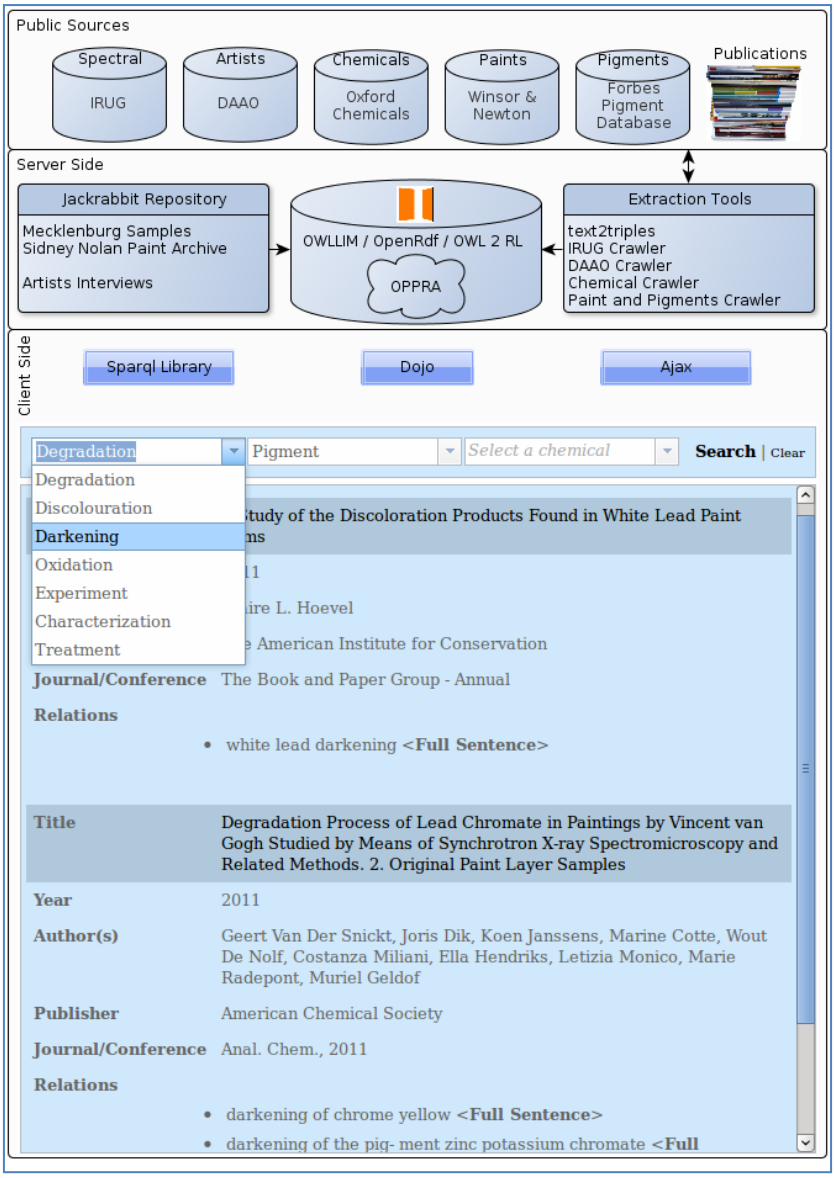



**Figure 8.4: A high-level view of the DALI system architecture**

## 8.6. User Interface

To overcome the difficulties that non-technical users face in generating SPARQL queries, a user-friendly interface has been developed that automatically maps user input (via pull-down menus based on OPPRA terms) to SPARQL queries. The DALI user interface (20thcpaint, 2012a) enables conservators and scientists to seamlessly search for particular paintings, artists, paints, types of degradation, chemical compounds or characterisation methods across the integrated datasets.

The DALI server (based on the OWLIM (Bishop et al., 2012) implementation of the OpenRDF Sesame triple store) provides two ways to search the OPPRA-based knowledge-base. These are:

- The Data Aggregation and Linking Interface (HTTP/GET) that provides both keyword-based and advanced search options. This approach maps users' input to SPARQL queries, and returns a set of matching results that are displayed. Figure 8.5 shows a screenshot of a search via the keyword '*darkening*' – with a SPARQL screenshot constructed automatically based on the given keyword. The keyword-based search accepts a string representing a synonym that is mapped to a given concept (e.g., *?concept oppra:hasSynonym "synonym"*). The advanced search accepts a list of strings – each representing a synonym that is restricted to a particular concept in the OPPRA ontology (e.g., publication, author, painting, characterisation technique and treatment activity);
- REST/SPARQL that accepts a *query* type (e.g., SPARQL – *select, construct,* and *ask*) and *output* format (e.g., *application/rdf+xml, text/rdf+n3* and *application/x-trig*), and returns the appropriate graph (or collection of graphs) based on the given query type and output format.

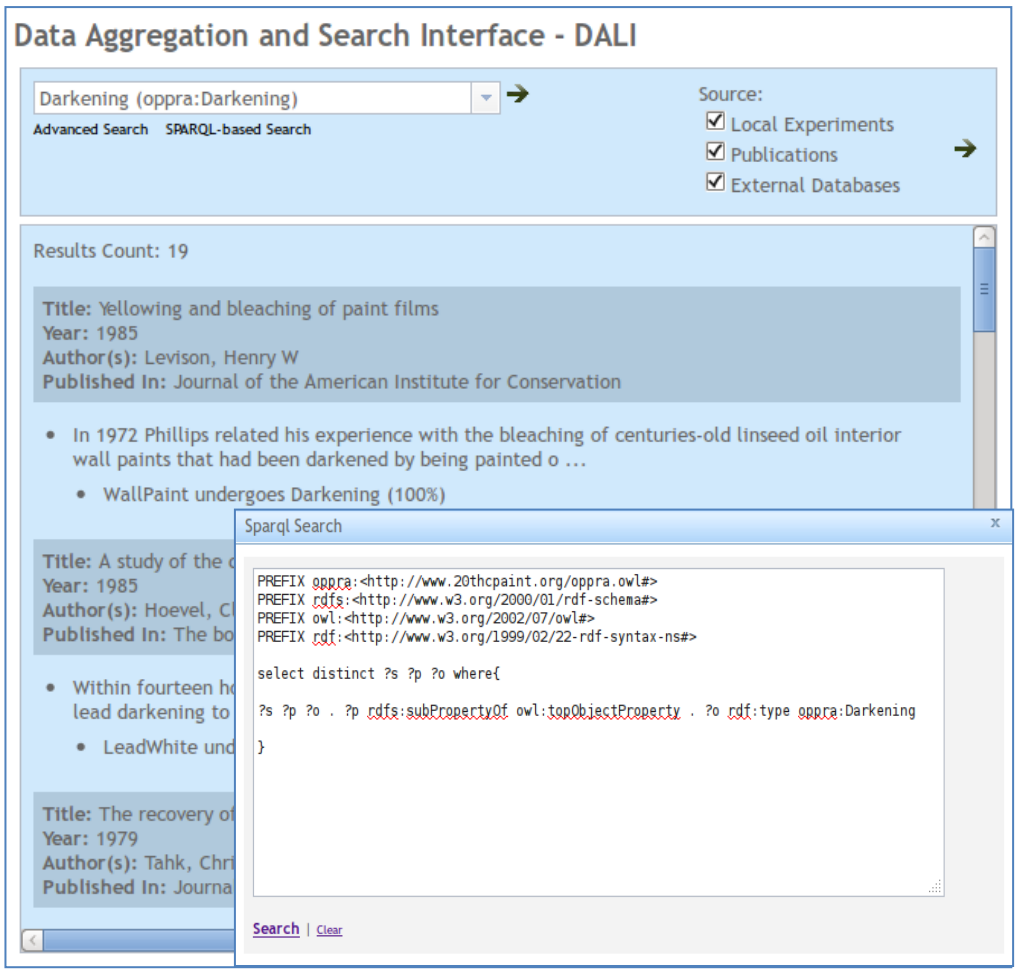



**Figure 8.5: User interface for a keyword-based search – mapped to a SPARQL query using DALI**

End users can also restrict the datasets that they wish to search (see top left hand side of Figure 8.5). For example, they may choose to search one or more of: local experimental data, publication data or external databases.

The following result types are currently supported in the implemented framework:

- Responses to sophisticated queries that involve multi-disciplinary domains (art history and materials science) with inferencing. For example, the following SPARQL statement returns the solvents that remove the varnish layer used in the painting *Epiphany*:

```
PREFIX oppra:<http://www.20thcpaint.org/oppra.owl#>

select distinct ?solvent where{
 ?artist oppra:paints oppra:Painting_Epiphany ;
         oppra:performedPaintingProcess ?paintingProcess .
 ?paintingProcess oppra:usedMaterial ?varnish .
 ?varnish oppra:wasRemovedBy ?solvent
}
```

- Finding sources (*oai-ore:Aggregation*) based on a given query. For example, the following SPARQL statement returns the graphs '*oppra:Publication*' that provide information on the SEM characterisations of samples taken from Sidney Nolan paintings:

```
PREFIX oppra:<http://www.20thcpaint.org/oppra.owl#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>

select distinct ?publication where{
  graph ?publication{
    ?sample oppra:takenFrom ?painting .
    ?painting oppra:paintedByoppra:SidneyNolan .
    ?sample oppra:characterisedBy ?characterisation .
    ?characterisation oppra:usedCharacterisationTechnique
oppra:ScanningElectronMicroscopy
  } . ?publication rdf:typeoppra:Publication
}
```

The user interface also enables the visualisation of results via the visualisation of RDF triples/graphs, as seen in Chapter 6. Figure 6.9 (in Chapter 6) shows an example of RDF triples extracted from a Sidney Nolan Archive record 'SampleRecord6' (Ripolin Paint, Black No. 1105) which was also a result of the query: *list records/publications illustrating FTIR, and Py-Gc-Ms characterisations of samples taken from Ripolin*.

## 8.7. Evaluation Results

This section provides details on the results of evaluating the capabilities of the knowledge-base and the underlying OPPRA ontology. The measures used for this assessment are: the precision of the information retrieval, the precision of the knowledge-base, and usability of DALI.

### 8.7.1. Precision of the Information Retrieval

This evaluation is performed by calculating the precision of the document or segment retrieval based on given SPARQL queries (i.e., retrieving references/experiments, and pin-pointing the exact sentences/triples that answer each query). The Precision (*P*) is defined in the experiments as the ratio between the correctly retrieved results (documents or segments), and the number of retrieved results as follows.

$$P_{\text{doc}} = \frac{Number\ of\ Correctly\ Retrieved\ Documents}{Number\ of\ Retrieved\ Documents}$$

$$P_{\text{seg}} = \frac{Number\ of\ Correctly\ Retrieved\ Segments}{Number\ of\ Retrieved\ Segments}$$

Based on the requirements and feedback of the 20th Century in Paint team, three sets of query types were identified and used in the evaluation: condition changes of materials, and/or their mention in online publications; investigations of materials and their degradation mechanisms; and multi-disciplinary questions involving the art history and conservation science domains. Chapter 3 (Section 3.5) provides examples of these queries. The results of each query are assessed by manually calculating the precision.

Queries are performed against the integrated OWLIM knowledge-base that contains the data extracted from datasets defined in Section 8.3.3 (experimental data, structured data from publications, external databases, and inferencing).

The integrated data was indexed using the following three indexing methods:

- Full reference indexing that uses Apache Solr/Lucene full-text indexing (Apache, 2011) for all documents and experiments – given the full text of each document and experiment as a <content>markup to be indexed;

- Breaking each document and experiment into smaller documents of three sentences (and sub-sections in case of experiments), and applying Apache Solr/Lucene full-text indexing to the sentences/sub-sections;
- Breaking each document and experiment into smaller documents of one sentence (and sub-section in case of experiment), and applying Apache Solr/Lucene full-text indexing to the sentence/sub-sections.

DALI only involves a sentence and an experiment (record)-based indexing since references (publications and experiments) are retrieved using the OAI-ORE model (e.g., Reference *oai-ore:aggregates* Sentence . Sentence *oai-ore:aggregates* Triple).

There are two reasons for using windows of one and three sentences and sub-sections in addition to the full-text indexing. These are:

1. To determine whether breaking a document/experiment into smaller parts (small graphs) for indexing achieves the same accuracy as indexing the entire document/experiment (overall graph);
2. To narrow down the segment retrieval (i.e., to precisely pin-point the triple(s) responsible for answering each given query).

Three different techniques for searching the knowledge-base are compared. These are:

- A keyword-based search using the REST-like API of Solr. This method converts queries to Solr queries. Each Solr query consists of a keyword pair with an "*AND*" operator. For example, the query "*find references that report on the cleaning of blistered oil paints with mineral spirits*" would be converted to the following three Solr queries:
  - "*blistered AND oil paint*";
  - "*cleaning AND oil paint*";
  - "*cleaning AND mineral spirit*".
- A thesauri-based search using the REST-like API of Solr as above, but with the addition of further queries for all of the keywords' synonyms (and the sub-classes' synonyms if any) from the OPPRA ontology (e.g., *oppra:hasSynonym*, and *rdfs:subClassOf*). For example, to search for references that mention "*the discolouration of synthetic resin paints*", the following queries are constructed:
  - *discolouration AND synthetic resin paint;*

- *discoloration AND acrylic paint;*
- *darkening AND plastic paint;*
- *fading AND polymer paint;*
- *staining AND vinyl paint;*
- *etc.*

- An ontology-based search using the SPARQL API of DALI. Using this method, queries are constructed using the SPARQL *select* query as explained above in Section 8.5.

Figure 8.6 illustrates the precision results for each query method graphically. Figure 8.6 shows that full reference indexing generally achieves better document retrieval than window indexing (keyword-based search: 44% (full reference) vs. 31% (3 sentences) vs. 22% (1 sentence) / thesauri-based search: 69% (full reference) vs. 56% (3 sentences) vs. 43% (1 sentence)). This is because it supports searching across the overall graph (full reference) that includes the keywords of interest (e.g., *darkening*, *blistered*, and *oil paint*), rather than searching smaller parts of each document (small graphs that would not be likely to include all the keywords of interest (e.g., a sentence/sub-section that includes the terms *darkening*, *blistered*, and *oil paint*)).

On the other hand, segment retrieval using full reference indexing usually has lower precision than window indexing (keyword-based search: 6% (full reference) vs. 13% (3 sentences) vs. 21% (1 sentence) / thesauri-based search: 18% (full reference) vs. 24% (3 sentences) vs. 35% (1 sentence)) – because of the granularity of the indexing process within these windows (full reference vs. three sentences/sub-sections vs. one sentence/sub-section).

Although DALI involves sentence-based indexing, it achieves better results on both document/experiment retrieval and segment retrieval tasks (89% and 87% respectively) than on a keyword-based search (best values of 44% (document) and 21% (segment)) and thesauri-based search (best values of 69% (document) and 35% (segment)). The high precision of document and segment retrieval can be justified by the following:

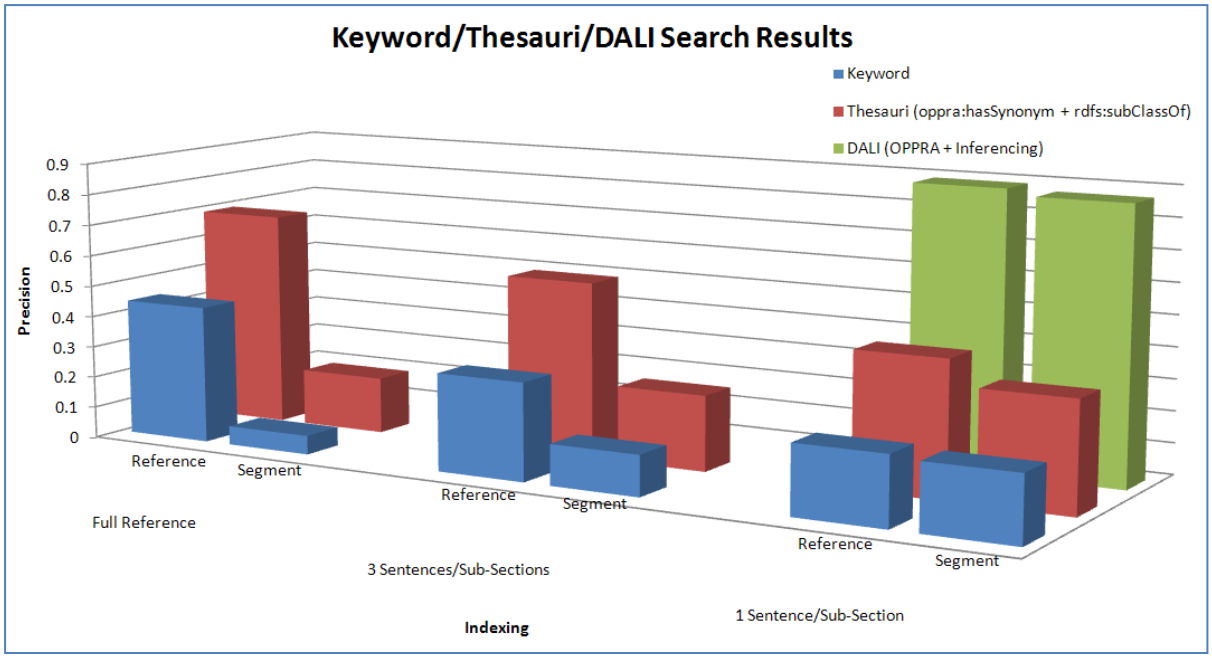



**Figure 8.6: Reference/segment retrieval precision based on keyword, thesauri, and DALI search of three indexing methods (full reference, 3 sentences/sub-sections, and 1 sentence/sub-section)**

- The indexing of triples is determined by how the knowledge is stored (i.e., by its structure which conforms to the OPPRA ontology). For example, the extracted triples from each sentence/sub-section are stored as OPPRA-conformant triples (e.g., OilPaint *undergoes* Darkening), but are aggregated using the OAI-ORE model that includes the transitive property "oai-ore:*aggregates*". This property acts as a Named Graph-based storage of each extracted triple, but it also involves transitive inferencing of the graphs containing these triples (e.g., Reference oai-ore:*aggregates* Sentence; and Sentence oai-ore:*aggregates* Triple).
- If a core triple (or multiple core triples) correctly matches the query, the aggregated references (e.g., Reference) will correctly match the same query via inferencing. This is due to the fact that knowledge (references) is not derived from different indexing methods (as in the case of Solr indexing). For example, the following three aggregations will result in the fourth aggregation (by inferencing):
  - Sentence5 *aggregates* (OilPaint *undergoes* Darkening), (Heat *causes* Darkening);
  - Sentence7 *aggregates* (Light *causes* Recovery), (Recovery *occursIn* OilPaint);
  - Reference9 *aggregates* Sentence5, Sentence7;

  - → Reference9 aggregates (OilPaint *undergoes* Darkening), (Heat *causes* Darkening), (Light *causes* Recovery), (Recovery *occursIn* OilPaint) – these aggregations are not derived from different indexing methods.
- Inferencing is a major contributor to the high retrieval accuracy. The inferencing rules applied to the knowledge-base (OPPRA-based indexed corpus) are:
  - Transitive properties: aggregates, consistsOf;
  - consistsOf ← wasSampleSource . containsMaterial;
  - materialFormsPartOf ← materialFormsPartOf . takenFrom;
  - paintedOn ← wasPaintedBy . hasTimespan;
  - isPaintingDateOf ← isTimespanOf . producedPainting;
  - hasTechnique ← performed . usedSpecificTechnique;
  - techniqueOf ← techniqueWasUsedBy . carriedOutBy;
  - isReferredToBy ← undergoes . outputs;
  - refersTo ← isOutputFrom . concerned;
  - painted ← performedPaintingProcess . producedPainting;
  - paintedBy ← wasPaintedBy . performedByArtist;
  - paintedWithTechnique ← wasPaintedBy . usedArtisticTechnique;
  - indicates ← indicates . consistsOf;
  - undergoes ← materialFormsPartOf . undergoes.

### 8.7.2. Precision of the Knowledge-Base

The precision of the knowledge-base (OPPRA-based RDF triples) is measured by comparing its facts (instances and relations between these instances) to the ground truth. Since the ground truth is difficult to obtain (because OPPRA's facts are generated via three different services – experimental result capture, machine learning extraction tools, harvesting of information from external datasets), manual measurements are calculated as follows:

- Random selection of a number of facts (1000 facts) that exist in the knowledge-base;
- The marking of each fact as correct or incorrect – by inspecting the source where the fact appears (e.g., publication or database);
- Calculation of the precision (*P*) as follows:

$$P = \frac{number\ of\ correct\ facts}{total\ number\ of\ facts\ inspected}$$

Table 8.1 shows the precision results for the facts in the OPPRA-based knowledge-base. The evaluation indicates high precision results with the following observations:

- Facts generated via extraction from structured databases (internal and external), and inferencing yield a 97.4-99% accuracy – due to the manual cost associated with implementing the data capture services (e.g., internal repositories, extraction tools from external databases, and OWL 2 RL rules);
- Facts extracted from the *text2triples* platform yield an accuracy of 85% – due to the automatic extraction services that were presented in Chapter 7.

**Table 8.1: Results for the OPPRA-based knowledge-base precision**

| Source | Number of Facts | Precision |
|---|---|---|
| Internal Databases | 200 | 98.8% |
| *text2triples* | 500 | 85% |
| External Databases | 200 | 97.4% |
| Inferencing | 100 | 99% |
| **Total** | **1000** | **95.05%** |

Since the facts are defined, verified and recorded by conservators and scientists, the correctness of the facts is not measured according to their validity in the real world, but on the extent to which they preserve the same meaning as the original source. Inferred facts are assessed based on the correctness of the application of OWL 2 RL rules (*input*: facts from multiple data sources, and *output*: new facts).

### 8.7.3. Usability of DALI

The usability of the DALI search system was evaluated by deploying it within the team of the 20th Century in Paint collaborators (6-8 art conservators and materials scientists) who are investigating different case studies within the *20th Century in Paint* project. The usability testing was performed via hands-on demos, joint use and documentation of feedback conducted during meetings with the *20th Century in Paint* project team. In addition, the project team as well as (2-3) user interface experts from the School of ITEE at the University of Queensland had online access to the prototype system (20thcpaint, 2012a), so could provide continual feedback (via an

iterative testing and refactoring process) to both the system's front-end user interface and functionality.

Based on the usability testing conducted to-date, the collaborators' feedback indicated that the use of DALI greatly increased the speed and efficiency at which users could search, aggregate, and analyse art/paint conservation data and information. The iterative and refactoring procedures to DALI achieved the following outcomes:

- Simplified querying: queries are formulated through Web-based graphical interfaces that search across key art history and materials science databases. By typing/selecting terms in the OPPRA-based auto-complete text fields, a user can pose complex queries without having to understand or synthesise different terminologies or having to navigate through different search interfaces.
- Improved efficiency, and optimisation of queries by:
    - Storing heterogeneous data from disparate datasets locally into a robust, structured OWL ontology (OPPRA) for art/paint preservation;
    - Formulating hypotheses from the knowledge acquired – i.e., reasoning over the aggregated data using OWL 2 RL;
    - Eliminating various problems such as network bottlenecks, low response times, and the unavailability of sources.
- Enhanced flexibility - conservators and materials scientists can choose which data sets to include in their searches so can tailor the search interface to only use the datasets of relevance to their interests or only use those they trust.
- Improved provenance – the search results include provenance information that includes the original source of RDF facts as well as visualizations that show the provenance of inferred facts.

The usability testing, however, also revealed the following limitations:

- Currently there is no ranking of search results. Ideally the most relevant matching results should be ranked at the top of the results.
- Certain expert users should be permitted to view and modify/correct records in the knowledge-base that are incorrect. In order to actually proceed with the modification, users need to separately open the required user interface (e.g., Experimental Data Capture, *text2triples*), login, and perform these modifications.

- There is currently no interface for users to enter, view, and edit inferencing rules – this would be a very useful addition to DALI that would capture domain expert knowledge so it can be re-used and refined over time as the domain expert knowledge grows and improves.
- There are currently a number of relevant and valuable sources of information that are lacking from the knowledge-base due to access restrictions. For example, detailed provenance information about individual artworks is difficult to acquire and remains sensitive, confidential information held within many art gallery databases. Hopefully over time, public cultural institutions and art galleries may become more open with such information. Similarly there are a number of commercial databases associated with artists' paints and paint materials that contain valuable data, but they were outside the scope and budget of this project.

## 8.8. Summary

This chapter presents a Semantic Web approach to data integration for 20th century art/paint conservation through the development of the DALI system. DALI leverages and integrates a variety of services developed for the *20th Century in Paint* project (e.g., OPPRA, structured data capture and extraction from local databases, external databases, and publications) to describe, integrate and infer information for the art/paint conservation community.

The semantic search functionality and the OWL 2 inferencing provided through DALI produced encouraging results that indicate that the approach adopted within DALI has enabled enhanced and integrated access to cross-disciplinary information for the art conservation community. DALI provides answers to more sophisticated queries than traditional data integration tools – e.g., integrating, re-using and reasoning across datasets from distributed sources.

However this research also highlighted a number of areas that require further research including:

- Improving automatic entity resolution, and automatically identifying similarities and relationships between different URIs in the knowledge-base;

- Identifying additional emerging, relevant data sources that provide information on art/paint conservation, and automating the extraction of new or updated data, and its incorporation into the OPPRA-based knowledge-base;
- Implementing a full system integration that allows users to add, update, access, search and share data without navigating in and out of the various services used for the 20th Century in Paint project – Experimental Data Capture, *text2triples,* DALI, and visualisation tools;
- Obtaining user feedback to DALI by conducting usability studies with a wider community of users e.g., the APTCAARN community in Asia-Pacific or the International Network for the Conservation of Contemporary Art (INCCA).

# Chapter 9

# Conclusions and Future Work

## 9.1. Summary of the Research

As stated in Chapter 1, the general research question that has been addressed in this thesis is: “Can a collaborative distributed knowledge-base and decision support platform be built to help answer sophisticated questions about art/paint conservation?” This question can be broken down into the following twelve, more specific, research questions:

1. Can a comprehensive knowledge-base comprising RDF graphs be built to support art conservators’ information requirements?
2. What is the quality of the data model – including the upper ontology, provenance ontology and other ontologies – for underpinning the knowledge-base?
3. What sub-disciplinary ontologies exist or need to be developed and incorporated?
4. Do existing data models (e.g., CIDOC-CRM) support the requirements of this project or do they need to be extended or refined?
5. Is there an existing ontology for describing art deterioration, preservation and conservation concepts?
6. If not, are there existing controlled vocabularies that can be re-used to describe artists’ materials, paints, painting terminology, conservation terminology, preservation terminology (e.g., techniques, materials and instruments)?
7. Can experimental data (samples, experimental processes, observations/measurements, characterisations) be captured and stored in a standardised machine-processable format?

8. How accurate is the structured knowledge (that conforms to the ontology, and that is extracted from relevant publications, to enable the re-use, integration and comparison of emerging, current and past knowledge)?
9. How efficient and accurate can a large corpus of RDF graphs (derived from publications, related databases and experimental data) be for aggregating, searching, browsing and retrieving (via SPARQL) conservators' information?
10. Can semantic inferencing and reasoning (e.g., OWL-DL) be enabled across the RDF graphs in order to extract previously unknown knowledge?
11. Can publications about art conservation be linked to raw and derived experimental datasets using RDF graphs?
12. How can the improvements and benefits of such data models and services for the art conservation community be evaluated?

These questions have been formulated based on the review of related work described in Chapter 2 and are the motivation for the following principal objectives/outcomes:

- The design and development of the OPPRA ontology;
- The design and development of a knowledge-base to support the storage of experimental data, structured data (extracted from publications) and external databases;
- The design and development of a collaborative experimental data repository;
- The development of text mining tools to extract structured knowledge from past publications;
- The development of a SPARQL search interface to provide access to the distributed, heterogeneous knowledge captured (via the experimental data capture, text analysis, data capture from the external databases and semantic inferencing) for the art/paint conservation domain.

In the following sections, the contributions with respect to these aspects are summarised. The potential areas for further investigation are then discussed.

## 9.2. Main Original Contributions

Based on the research questions outlined in Chapter 1 (Section 1.5), this research makes the following five original contributions to the field of cultural heritage

informatics: the OPPRA ontology, the OPPRA-based knowledge-base, the collaborative experimental data capture, the automatic knowledge extraction tools, and the data aggregation, linking and querying interface.

### 9.2.1. The OPPRA Ontology

Chapters 4 and 8 address Research Questions 2-6 and 12. They described the first contribution of this dissertation, namely:

- The first ontology (OPPRA) to support the information integration and analysis requirements of art/paint conservators.

The OPPRA ontology has been developed to support the information integration and analysis requirements of art conservators, and to underpin the knowledge-base (comprising the OWL model, OWL 2 RL rules, and instances captured/extracted from experiments, publications and external data sources). There has been no previous attempt to develop an ontology that defines the entities and attributes associated with paint (its chemistry, composition, additives, and behaviour), its degradation over time (chemical reactions), and the effect of environmental parameters. No previous ontology has attempted to link materials science with analytical and art conservation techniques.

The OPPRA ontology satisfied the objectives of streamlining the requirements of the art/paint conservation community. In the context of the 20$^{th}$ Century in Paint project, for example, OPPRA was successfully used to: 1) document and describe experiments conducted by the art/paint conservators; 2) automatically extract structured data about past research and experiments from relevant publications; and 3) bridge the gap between the physical and digital provenance of paintings and paint samples. OPPRA fulfilled these functions by providing a common, machine-readable formal representation of the knowledge in the domain of art/paint preservation.

### 9.2.2. The OPPRA-based Knowledge-base

Chapters 5 and 8 concerned Research Questions 1, 9-10 and 12. They described the second contribution of this dissertation, namely:

- The OPPRA-based knowledge-base to support the storage of experimental data, structured data (extracted from publications) and external databases – required for informed decision-making by the art/paint conservation community.

An OPPRA-based knowledge-base has been established to support the storage of experimental data, structured data (extracted from publications) and external databases – as required for informed decision-making by the art/paint conservation community. No previous research has attempted to use semantic formalism to integrate data associated with paint composition, paint processes (including chemical processes and degradation over time), the effect of environmental parameters on paint or the effect of different conservation treatments on paint. No knowledge-base exists to provide semantic information on art/paint conservation, in a form that facilitates its discovery, re-use and aggregation.

The OPPRA-based knowledge-base satisfied the requirements of the conservators and scientists involved in the 20$^{th}$Century in Paint project by delivering a set of services which are simple, flexible, intuitive and efficient. For example, the knowledge-base was successfully used to underpin the DALI search interface that aggregates information from internal and external datasets, and reasons across this information, to answer advanced and semantically linked queries such as: *What solvents will remove surface varnish from the painting Epiphany?*

### 9.2.3 Collaborative Experimental Data Capture

Chapter 6 concerned Research Questions 7, 11 and 12. It described the third contribution of this thesis, namely:

- A framework and set of services to support the capture, publishing, linking and searching of experimental data associated with art/paint conservation (based on the OPPRA ontology).

A framework and set of services has been developed to support the capture, publishing, linking and searching of experimental data associated with art/paint conservation. Previous efforts have focused on capturing scientific experiments in fields that include the biological sciences (Abidi et al., 2012, Smith et al., 2011), and chemical sciences (Krafft et al., 2010, Pirró et al., 2010, Reid and Edwards, 2009).

No previous work has focussed on capturing experimental data in the field of art/paint conservation. A major factor, which makes this applied research different from the other approaches, is the cross-disciplinary challenges associated with the art/paint preservation domain. This component needed to record the semantics of both the provenance of the paint samples (e.g., painting, paint, artist, genre) as well as the key concepts associated with art/paint chemistry (composition and materials) and characterisation and experiments that simulate deterioration mechanisms and the effects of different alternative treatments.

The experimental workflow system satisfied the functional and research requirements of the conservators and scientists involved in the 20$^{th}$ Century in Paint project. For example, the system was successfully used to enable conservators and scientists to: 1) create/define new projects; 2) add/remove team members 3) define sets of activities/tasks and inputs/outputs; 4) describe samples and associated characterisation images/data; 5) collaboratively edit/add data and observations; 6) attach access policies; 7) visualise and compare results; 8) access, share and re-use experimental results via persistent links (URLs); and 9) insert links from publications to experiments/experimental data in the knowledge base via persistent URLs.

### 9.2.4 Automatic Knowledge Extraction Tools

Chapter 7 addressed Research Questions 8 and 12. It described the fourth contribution of this thesis, namely:

- A set of text analysis tools (a GATE pipeline comprising NER and RE tasks) to support the extraction of structured data from publications about art/paint conservation (based on the OPPRA ontology).

A set of text analysis tools (a GATE pipeline comprising NER and RE tasks) was developed for extracting structured data from publications about paint conservation. RDF graph visualisation and editing tools were also developed to improve the accuracy of the extracted RDF (structured data). The text analysis approach used in this study differs from other approaches (in the chemistry (Na et al., 2010, Yamashita et al., 2011), and cultural heritage informatics domains (Byrne, 2009, Commetric, 2013)) because it is the first to focus on the specific requirements associated with art/paint preservation (by building the tools on the OPPRA ontology). This research

component extends and applies existing NER and RE techniques to extract structured knowledge about art/paint conservation from a publication corpus and represent it in OPPRA-compliant RDF – to enable comparison and integration of knowledge and facts embedded in full-text documents. This research also advances the current state of the art of NER and RE services by implementing an efficient Web-based text tagging system that allows users to define (and modify) named entities in text documents, and describe (and visualise) the relations between these entities based on an underlying data model (e.g., OPPRA).

The text analysis tools satisfied the functional and research requirements of the conservators and scientists in the 20th Century in Paint project. For example, these tools were successfully used to: 1) enable conservators and scientists to discover and re-use knowledge hidden within art conservation publications; 2) to visualise, edit/correct and link RDF triples extracted from publications; and 3) enable publications to be linked to experiments via URIs that point to Named Graphs.

### 9.2.5 Data Aggregation, Linking, and Querying Interface

Chapter 8 concerned Research Questions 1, 9-10 and 12. It described the fifth contribution of this thesis, namely:

- An interface (comprising OWL 2 RL inferencing, SPARQL search, and visualisation) to provide responses to complex cross-disciplinary queries about art/paint conservation, by integrating (and reasoning across) data from relevant existing databases, experimental datasets and publications.

The DALI framework (comprising OWL 2 RL inferencing, SPARQL search and provenance visualisation) provides responses to complex queries about art conservation and materials science, by integrating data from relevant existing databases, experimental datasets and publications. A major factor, which makes the data integration used in this study different from other approaches in the cultural heritage informatics domain (Aliaga et al., 2011, Hyvönen et al., 2009, Binding, 2010, Binding et al., 2008, Hyvönen et al., 2006, Mellon, 2009, Monroy et al., 2010, Toledo et al., 2009), is the focus on the specific requirements and cross-disciplinary concepts associated with art/paint preservation. DALI integrates diverse databases about the provenance of paintings (collection, exhibition, condition assessment, and

treatment), artists' techniques (artist, period, genre, source of materials, additives, techniques), paint composition (pigments and paint formulation databases) and materials science (physical and chemical properties, analytical techniques (SEM, TEM, Infrared multispectral techniques, Raman microscopy, X-Ray diffraction), and characterisation data), and applies reasoning over the aggregated data to help art/paint conservators answer the central questions of their studies (e.g., Under what conditions do metal soaps form? What are the causes of metal soap formation, aggregation and extrusion? How should metal soap extrusion be treated?).

DALI satisfied the functional and research requirements of the conservators and scientists in the 20th Century in Paint project. For example, the interface was successfully used to: 1) provide a single Web-based search interface to: the 20th Century in Paint project databases (Sidney Nolan Paint Archive, and Mecklenburg Samples); structured data extracted from past publications via the *text2triples* software; and a set of related publicly available databases (e.g., W&N, DAAO, IRUG Spectral Database, and CAMEO); and 2) answer sophisticated and multi-disciplinary queries about art/paint conservation that were not previously possible through a single search interface.

### 9.2.6 Original Technical Contributions Independent of the Art Conservation Application

The novelty of the work described here does not rest solely on the uniqueness of the application domain (i.e., art conservation). Original technical contributions that are independent of the art conservation application include:

- The Experimental Data Capture system – a Web-based collaborative system that enables scientific teams to describe their activities (e.g., experiments), and share experimental results using role-based access controls (e.g., microscopic images, spectrographic/FTIR data, annotations). This system differs from the other available systems for describing and sharing experimental results by: 1) enabling scientific teams to link to similar experiments conducted by others; 2) enabling scientific teams to link experiments and experimental outputs to publications via named graphs.

- The text2triple system – a more efficient Web-based system for tagging text, and extracting structured data/knowledge from text publications, that incorporates the following GATE plugins:
  - An NER plugin that automatically tags (and allows users to modify and add new) named entities;
  - An ambiguity resolution plugin to extract ambiguously named entities;
  - An RE plugin that automatically finds (and allows users to modify and add new) relationships between named entities;

  This system differs from the other available systems for text processing by providing functionalists of GATE (e.g., machine-learning, and user-based support for text tagging, ambiguity resolution, synonyms suggestion using WordNet, and relation extraction) from within the the browser (i.e., no need for any software (or plugins) to be installed by the user).

- The DALI search engine – a Web-based system that integrates cross-disciplinary data from distributed databases (both local experimental data and publicly available databases) (and applies inferencing based on a pre-defined rules). The system enables users to seamlessly perform complex queries across multiple data sources and multiple disciplines (art history and chemistry) via the back-end ontology. This approach differs from the other available search engines by providing higher precision of both document and segment retrieval (i.e., accurate referencing as to where (in the document/record) the search results match the query).

## 9.3. Limitations, Future Work and Open Challenges

This section discusses the issues considered to be the main limitations of the work presented in this thesis, identifies the potential areas for further investigation and discusses open challenges.

### 9.3.1. Limitations of the Research Results

A number of limitations were identified within the specific implementations and research results produced within this thesis.

The OPPRA ontology is currently limited with regard to certain specific high-level concepts that are significant within the art conservation domain such as time and

temporal relations (e.g., Time ontology (Hobbs and Pan, 2006)),or place and spatial relations(e.g., Geospatial ontology (Lieberman et al., 2007)). It is believed that the OPPRA ontology is able to incorporate such additional ontologies through extensions, in the same way it incorporates the OreChem ontology.

Currently there is no interface that enables the art conservation community to interactively and collaboratively edit/refine the OPPRA ontology. Provision of an online easy-to-use collaborative editing interface, accessible to authenticated experts, would be the quickest and most efficient way to improve the ontology over time. Furthermore, investigating the best ontology library for publishing the OPPRA ontology to the Semantic Web and exposing the ontology to the art conservation community is worth pursuing.

The Experimental Data Capture component is limited in that it does not support importing data (experiments/sub-experiments) from other content management and experimental workflow systems (e.g., Kepler, Taverna, and myExperiment Virtual Research Environment). In the future, an '*import*' option/functionality should be developed to allow researchers to easily incorporate their previously conducted experiments into the Experimental Data Capture framework, and in turn publish them into the OPPRA-based knowledge-base. In addition, the Experimental Data Capture component is limited in that it does not describe indicative conditional branch statements (i.e., logical operations that act upon OPPRA's entities and govern the experimental workflow/process). Finally, being able to capture and share the workflow patterns associated with experiments would enable greater comparison and re-use of experimental data.

Although the *text2triples* framework and structured data extraction process achieved satisfactory results (within the NER and RE tasks), there remains room for improvement in performance and accuracy. The speed and efficiency of the OPPRA-based gazetteer, for example, can be improved by caching the OPPRA ontology while the GATE resources are being loaded, and documents are being opened. Furthermore, the efficiency and accuracy of the NER and RE tasks can be improved by pre-processing the full publication before giving all of the sentences (MALLET instances) to the NER and RE classifiers (e.g., reducing unnecessary sentence inputs by segmenting the publication into titles, sections, figures, tables, references

and footnotes). Finally, the precision of the OPPRA-based gazetteer is anticipated to improve as the corpus of publications (tagged with named entities and relations) expands, and the OPPRA ontology becomes more complete and accurate.

The DALI framework is limited in that it currently separates the search interface from the data ingestion frameworks (*text2triples* and Experimental Data Capture). The search results provide links to the records that users are allowed to view/modify, but in order to actually proceed with the modification, users need to open the required editing software separately, login, and perform the modifications. In the future, an integrated framework should be developed that allows users to add, update, access, search and share data without navigating in and out of the various services used for the 20$^{th}$Century in Paint project. DALI could also be improved by enabling users to specify individual projects and or external databases that they want to include/exclude in searches.

The OPPRA-based knowledge-base is currently incomplete and lacks comprehensive coverage of data across many topics. A large number of relevant databases and valuable sources of data are inaccessible due to access restrictions. For example, detailed provenance information about individual artworks is difficult to acquire and remains sensitive, confidential information held within many art gallery databases. Hopefully over time, public cultural institutions and art galleries will adopt a more “open access” approach to data. Similarly, there exist a number of commercial databases that contain valuable data but the cost of a license and licensing restrictions prohibit wide accessibility or incorporation within systems such as DALI.

Finally, the adoption of external data indexing/mapping tools (e.g., Web crawling, and D2R for database/RDF mapping)is expected to result in unreliable and possibly outdated results. Automated harvesting services that regularly check for new or updated data within external databases, and then reflect those modifications in the OPPRA-based knowledge-base are needed.

### 9.3.2. Future Research Directions

Future work plans for the knowledge-base and associated tools and services include:

- Implementing an interface to enable users to enter, view, and edit inferencing rules. This interface would be a very useful addition that would capture domain expert knowledge so it can be re-used and refined over time as the domain expert knowledge grows and improves;
- Investigating if languages other than English can be incorporated to serve the multi-lingual art/paint conservation community. The current system only supports English. Further research would be required to determine: how much work would be required to support the documentation, querying and reasoning over facts recorded in other languages? And which system components will have to be modified and/or extended?
- Evaluating the SPARQL-based search interface in order to determine if it provides better query performance and improved precision and recall over traditional publication search engines;
- Providing ranking measures for search results. This functionality is currently missing; ideally the most relevant matching results should be ranked at the top of the results.

In addition, one aspect of the resulting framework and services that has not been fully evaluated is the scalability of the system. To date, the knowledge base contains 114969 triples (8790 explicit triples from data sources, and 106179 triples obtained from inferencing). Compared to biomedical databases for example, this is a relatively small knowledge-base. As it expands with time, the question is whether the current design will scale? Will large communities of users be able to execute queries and retrieve easily-interpreted responses in a reasonable time-frame (e.g., matter of seconds)? This issue will need to be monitored over time to determine if the SPARQL querying may need to be optimised.

Finally, carrying out a broader user evaluation and usability study of the system with the collaborators of the 20$^{th}$Century in Paint project will help to define a better user experience, and inform the development of further services to be offered. Carrying out a broader user evaluation and usability study of the system with collaborators

outside of the 20th Century in Paint project will also help to provide a better user experience, and inform the development of further services. For example, it would be advantageous to evaluate the system with the broader APTCAARN community (in Asia-Pacific) or with art conservation communities in Europe and the US (e.g., International Network for the Conservation of Contemporary Art (INCCA)).

### 9.3.3. Open Challenges – Applying Semantic Web Technologies to Art Conservation

The research in this thesis highlighted a number of unresolved issues/challenges which became apparent when applying Semantic Web technologies to the capture, re-use and reason over art conservation knowledge:

- Automatic entity resolution remains a very challenging task. For example, the tasks of identifying when two entities are the same (e.g., people, artists, paintings, pigments, chemical compounds, samples), and assigning them the same URI are essential for any informatics-enabled system, and for Linked Open Data generally. Despite good progress in entity resolution methods (Chieu and Teow, 2012, Kim et al., 2004, Krallinger et al., 2013, Liao and Zhang, 2012, Zhang and Elhadad, 2013), such approaches have not yet been implemented or optimized for the art conservation domain. Challenges arise, for example, due to variations in how an entity may be referenced (e.g., '*Sydney Nolan'*, and '*the artist'* (Nolan)), or from the existence of several entities with the same name (e.g., '*Sydney Nolan'*, and '*S. Nolan'*), or even from spelling mistakes in the name.
- Cultural and research organisations tend to be reluctant to share information. For examples art conservators tend not to publicise mistakes. Scientists want exclusive access to their data so they can be the first to publish new findings. Art galleries are also highly sensitive when it comes to copyright issues or information associated with the provenance of art works. Thus, issues associated with data ownership, permission of use, trust, and copyright need be addressed and resolved before initiatives like Linked Open Data are fully embraced by the agencies/communities involved in art conservation.
- It is relatively easy to build an online digital archive, but establishing an online community of enthusiastic researchers and scholars who frequently contribute high quality content and knowledge to an existing knowledge-base is a much

greater challenge. Establishing and maintaining an active online community of users is a significant social problem but one which is more easily overcome: if the underlying technologies are fast, simple, intuitive, collaborative and useful; if there is a dedicated community liaison person employed on outreach activities; and if the project employs social networking tools such as Facebook and Twitter to constantly engage with the community and highlight valuable contributions.

## 9.4. Summary

This thesis described the results of a collaboration with the 20th Century in Paint project that aimed to develop a set of services to enable the extraction, creation and storage of knowledge about paint conservation – in an online semantic knowledge-base, so that it can be discovered, shared, re-used and reasoned across, by the art/paint conservation community.

The main outcomes and original contributions to the field of cultural heritage informatics are: the OPPRA ontology that underpins the knowledge-base; a repository to support the capture, storage, search and retrieval of experimental and characterisation data; semi-automatic techniques to extract structured data from existing publications; and advanced search and query interfaces that enable researchers to seamlessly integrate distributed databases on artists, artistic techniques, paints, chemicals, and chemical processes. The outcome is a framework that enables paint conservators to share their knowledge and results, to improve their understanding of paint degradation processes, and to identify and document new methods for stabilising, protecting and repairing our valuable but vulnerable paintings.

# Bibliography


20THCPAINT. 2010a. *The Ontology of Paintings and PReservation of Art - OPPRA* [Online]. Available: http://www.20thcpaint.org/oppra-owl/ [Accessed July 2014].

20THCPAINT. 2010b. *The Twentieth Century in Paint website* [Online]. Available: http://www.20thcpaint.org/ [Accessed July 2014].

20THCPAINT. 2012a. *Data Aggregation and Linking Interface - DALI* [Online]. Available: http://www.20thcpaint.org/dali/ [Accessed July 2014].

20THCPAINT. 2012b. *Structured Data Extraction using text2triples* [Online]. Available: http://www.20thcpaint.org/text2triples/ [Accessed July 2014].

ABIDI, S. R., ABIDI, S. S., KWAN, M. & DANIYAL, A. 2012. An Ontology Framework for Modeling Ocean Data and E-Science Semantic Web Services. *International Journal of Advanced Computer Science,* 2 (8)**,** 280-286.

ABU, A., LIM, S. L. H., SIDHU, A. S. & DHILLON, S. K. 2013. Semantic representation of monogenean haptoral Bar image annotation. *BMC Bioinformatics,* 14**,** 48-57.

ACS. 2011. *Analytical Chemistry* [Online]. Available: http://pubs.acs.org/journal/ancham/ [Accessed July 2014].

ADAMS, N., CANNON, E. & MURRAY-RUST, P. 2009. ChemAxiom – An Ontological Framework for Chemistry in Science. *International Conference on Biomedical Ontology.* Nature Publishing Group, 1, 2. http://dx.doi.org/10.1038/npre.2009.3714.1.

ADDIS, M., HAFEEZ, S., PRIDEAUX, D., LOWE, R., LEWIS, P., MARTINEZ, K. & SINCLAIR, P. 2006. The eCHASE System for Cross-border Use of European Multimedia Cultural Heritage Content in Education and Publishing. *AXMEDIS: 2nd International Conference on Automated Production of Cross Content for Multi-Channel Distribution.* Leeds, UK. http://eprints.soton.ac.uk/264265/.

ADDIS, M., MARTINEZ, K., LEWIS, P., STEVENSON, J. & GIORGINI, F. 2005. New ways to search, navigate and use multimedia museum collections over the web. *Museums and the Web.* Vancouver, Canada. http://eprints.soton.ac.uk/260909/.

ADUNA. 1997. *OpenRDF - Sesame RDF Triple Store* [Online]. Available: http://www.openrdf.org/ [Accessed July 2014].

AICCM. 1973. *AICCM Bulletin* [Online]. Available: http://www.aiccm.org.au/index.php?option=com_content&view=section&id=3&Itemid=44 [Accessed July 2014].

AICCM. 1999. *Visual Glossary* [Online]. The Australian Institute for the Conservation of Cultural Material. Available: http://www.aiccm.org.au/index.php?option=com_content&view=article&id=1&Itemid=2 [Accessed July 2014].

ALEXIEV, V., MANOV, D., PARVANOVA, J. & PETROV, S. 2013. Large-scale Reasoning with a Complex Cultural Heritage Ontology (CIDOC CRM). http://ceur-ws.org/Vol-1117/paper8.pdf.

ALIAGA, D. G., BERTINO, E. & VALTOLINA, S. 2011. DECHO - a framework for the digital exploration of cultural heritage objects. *Journal on Computing and Cultural Heritage (JOCCH),* 3**,** 1-26.

ALTINTAS, I., BARNEY, O. & JAEGER-FRANK, E. 2006. Provenance collection support in the kepler scientific workflow system. *Provenance and annotation of data***,** 118-132. Springer Berlin Heidelberg.

APACHE. 2004. *Apache Jackrabbit* [Online]. Available: http://jackrabbit.apache.org/ [Accessed July 2014].

APACHE. 2011. *Apache Solr* [Online]. The Apache Software Foundation. Available: http://lucene.apache.org/solr/ [Accessed July 2014].

APTCCARN. 2010. *Asia-Pacific Twentieth Century Conservation Art Research Network (APTCCARN)* [Online]. Available: http://cultural-conservation.unimelb.edu.au/partners/international/aptccarn/ [Accessed July 2014].

AROYO, L., STASH, N., WANG, Y., GORGELS, P. & RUTLEDGE, L. 2007. CHIP Demonstrator: Semantics-Driven Recommendations and Museum Tour Generation. *The Semantic Web,* 4825**,** 879-886.

ASHINO, T. 2010. Materials Ontology: An Infrastructure for Exchanging Materials Information and Knowledge. *Data Science Journal,* 9**,** 54-61.

ASWANI, N. & GAIZAUSKAS, R. 2010. Developing Morphological Analysers for South Asian Languages: Experimenting with the Hindi and Gujarati Languages. *Proceedings of the 7th Language Resources and Evaluation Conference.* 811-815. La Valletta, Malta. http://hnk.ffzg.hr/bibl/lrec2010/pdf/616_Paper.pdf.

AUER, S., BIZER, C., KOBILAROV, G., LEHMANN, J., CYGANIAK, R. & IVES, Z. 2007. Dbpedia: A nucleus for a web of open data. *The Semantic Web,* 4825**,** 722-735. http://link.springer.com/chapter/10.1007/978-3-540-76298-0_52.

BAGLIONI, P., GIORGI, R. & CHEN, C.-C. 2003. Nanoparticle technology saves cultural relics: Potential for a multimedia digital library. *Online Proceedings of DELOS/NSF Workshop on Multimedia Contents in Digital Libraries, 2-3.*

BARAK, M., HERSCOVIZ, O., KABERMAN, Z. & DORI, Y. J. 2009. MOSAICA: A web-2.0 based system for the preservation and presentation of cultural heritage. *Comput. Educ.,* 53**,** 841-852.

BARBOSA-SILVA, A., SOLDATOS, T., MAGALHAES, I., PAVLOPOULOS, G., FONTAINE, J.-F., ANDRADE-NAVARRO, M., SCHNEIDER, R. & ORTEGA, J. M. 2010. LAITOR - Literature Assistant for Identification of Terms co-Occurrences and Relationships. *BMC Bioinformatics,* 11 (1)**,** 70. http://www.biomedcentral.com/1471-2105/11/70.

BARUZZO, A., CASOTO, P., CHALLAPALLI, P., DATTOLO, A., PUDOTA, N. & TASSO, C. 2008. An intelligent service oriented approach for improving information access in cultural heritage. *In IACH'08: Information Access in Cultural Heritage (IACH) Workshop, European Conference on Digital Libraries, 299-304.*

BECHHOFER, S., HARMELEN, F. V., HENDLER, J., HORROCKS, I., MCGUINNESS, D. L., PATEL-SCHNEIDER, P. F. & STEIN, L. A. 2004. OWL - Web Ontology Language Reference. Available: http://www.w3.org/TR/owl-ref [Accessed July 2014].

BECKETT, D. & MCBRIDE, B. 2004. *RDF/XML Syntax Specification (Revised)* [Online]. W3C. Available: http://www.w3.org/TR/REC-rdf-syntax/ [Accessed July 2014].

BELMONTE, N. G. 2013. *JavaScript InfoVis Toolkit* [Online]. Available: http://philogb.github.io/jit/ [Accessed July 2014].

BERNERS-LEE, T. 2009. *Linked Data* [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html [Accessed July 2014].

BERNERS-LEE, T., FIELDING, R. & MASINTER, L. 2005. Uniform Resource Identifier (URI): Generic Syntax. IETF RFP 3986 (standards track), Internet Eng. Task Force. http://tools.ietf.org/html/rfc3986.

BERNERS-LEE, T., HENDLER, J. & LASSILA, O. 2001. The Semantic Web. *Scientific American,* 284 (5)**,** 28-37.

BILLINGE, S., RAJAN, K. & SINNOTT, S. 2006. From Cyberinfrastructure to Cyberdiscovery in Materials Science: Enhancing Outcomes in Materials Research, Education, and Outreach. *Report from NSF-sponsored workshop held in Arlington, Virginia.*

BINDING, C. 2010. Implementing Archaeological Time Periods Using CIDOC CRM and SKOS. *The Semantic Web: Research and Applications,* 6088**,** 273-287.

BINDING, C., MAY, K. & TUDHOPE, D. 2008. Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction Via the CIDOC CRM. *Research and Advanced Technology for Digital Libraries,* 5173**,** 280-290.

BISHOP, B., KIRYAKOV, A., TASHEV, Z., DAMOVA, M. & SIMOV, K. 2012. OWLIM Reasoning over FactForge. *Proceedings of OWL Reasoner Evaluation Workshop (ORE 2012), Collocated with IJCAR 2012.* 858, 1-147. Manchester, UK.

BIZER, C. & CYGANIAK, R. 2006. D2r server-publishing relational databases on the semantic web. *5th international Semantic Web conference,* http://richard.cyganiak.de/2008/papers/d2r-server-iswc2006.pdf.

BORKUM, M., LAGOZE, C., FREY, J. & COLES, S. 2010. A Semantic eScience Platform for Chemistry. *e-Science (e-Science), IEEE Sixth International Conference, 316-323. 7-10 December 2010, http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5689838.*

BRAY, T., PAOLI, J., SPERBERG-MCQUEEN, C., MALE, E., YERGEAU, F. & COWAN, J. 2006. XML 1.1 (Second Edition). W3C Recommendation, W3C. http://www.w3.org/TR/xml11/.

BUNDSCHUS, M., DEJORI, M., STETTER, M., TRESP, V. & KRIEGEL, H.-P. 2008. Extraction of semantic biomedical relations from text using conditional random fields. 9(1), 207. http://www.biomedcentral.com/1471-2105/9/207/.

BYRNE, K. 2009. Putting hybrid cultural data on the semantic web. *Journal of Digital Information (JoDI),* 10 (6), 1-22**,** http://homepages.inf.ed.ac.uk/kbyrne3/docs/jodi09kfb.pdf.

BYRNE, K. & KLEIN, E. 2010. Automatic extraction of archaeological events from text. *In: Proceedings of Computer Applications and Quantitative Methods in Archaeology.* Williamsburg, VA. www.academia.edu/download/30829923/caa09pres.pdf.

CARROLL, J., BIZER, C., HAYES, P. & STICKLER, P. 2005. Named graphs, provenance and trust. *Proceedings of the 14th international conference on World Wide Web.* 613-622. Chiba, Japan: ACM.

CHALLAPALLI, S., CIGNINI, M., COPPOLA, P. & OMERO, P. 2006. E-dvara: an xml based e-content platform. *AICA: Associazione Italiana per l'Informatica e il Calcolo Distribuito.*

CHEUNG, K., DRENNAN, J. & HUNTER, J. 2008. Towards an Ontology for Data-driven Discovery of New Materials. *AAAI Spring Symposium: Semantic Scientific Knowledge Integration, 9-14.*

CHIEU, H. L. & TEOW, L.-N. 2012. Combining local and non-local information with dual decomposition for named entity recognition from text. *Information Fusion (FUSION), 2012 15th International Conference on.* 231-238. IEEE.

CHIN. 2010. *Bibliographic Database of the Conservation Information Network (BCIN)* [Online]. Available: http://www.bcin.ca/ [Accessed July 2014].

COMMETRIC. 2013. *Commetric* [Online]. Available: http://www.commetric.com/Pages.aspx/Home [Accessed July 2014].

COOL. 2002. *Journal of the American Institute for Conservation* [Online]. Available: http://www.jstor.org/ [Accessed July 2014].

CORNEY, D. P. A., BUXTON, B. F., LANGDON, W. B. & JONES, D. T. 2004. BioRAT: extracting biological information from full-length papers. *Bioinformatics,* 20**,** 3206-3213.

CROFTS, N. 2004. Combining data sources–prototype applications developed for Geneva's department of historical sites and monuments based on the CIDOC CRM. Technical report, Direction du Patrimoine et des Sites, Geneva, 2004. www.cidoc-crm.org/docs/st_petersburg_combining_data_sources_.doc.

CROFTS, N., DOERR, M., GILL, T., STEAD, S. & STIFF, M. 2010. *Definition of the CIDOC Conceptual Reference Model (version 5.0.2)* [Online]. Available: http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.2.pdf [Accessed July 2014].

CROWLEY, R. S., CASTINE, M., MITCHELL, K., CHAVAN, G., MCSHERRY, T. & FELDMAN, M. 2010. caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. *Journal of the American Medical Informatics Association,* 17**,** 253-264.

CUNNINGHAM, H., MAYNARD, D. & BONTCHEVA, K. 2011. *Text Processing with GATE*, Gateway Press CA.

CUNNINGHAM, H., MAYNARD, D., BONTCHEVA, K. & TABLAN, V. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).* 168-175.

DAAO. 2010. *Dictionary of Australian Artists Online (DAAO)* [Online]. Design and Art Australia Online. Available: http://www.daao.org.au/main [Accessed July 2014].

DALE, R., MOISL, H. & SOMERS, H. 2000. Handbook of Natural Language Processing. *Computational Linguistics,* 27**,** 602-603.

DANICA, D., VALENTIN, T. & KALINA, B. 2008. A text-based query interface to owl ontologies. *In: LREC.* http://gate.ac.uk/sale/lrec2008/clone-ql/clone-ql-paper.pdf.

DE ROURE, D., GOBLE, C. & STEVENS, R. 2007. Designing the myexperiment virtual research environment for the social sharing of workflows. *e-Science and Grid Computing, IEEE International Conference on.* 603-610. IEEE.

DE ROURE, D., GOBLE, C. & STEVENS, R. 2009. The design and realisation of the myExperiment virtual research environment for social sharing of workflows. *Future Generation Computer Systems,* 25**,** 561-567.

DEGTYARENKO, K., MATOS, P. D., ENNIS, M., HASTINGS, J., ZBINDEN, M., MCNAUGHT, A., ALCÁNTARA, R., DARSOW, M., GUEDJ, M. & ASHBURNER, M. 2006. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res,* 36**,** D344–D350.

DOERR, M. 2003. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine,* 24**,** 75. http://www.aaai.org/ojs/index.php/aimagazine/article/view/1720.

DOJO-FOUNDATION. 2014. *Unbeatable JavaScript Tools - The Dojo Toolkit* [Online]. Available: http://dojotoolkit.org/ [Accessed July 2014].

DONKERSLOOT, W. 2009. *The Rembrandt Database Inter-institutional Research Resource on Paintings by Rembrandt* [Online]. Andrew W. Mellon Foundation. Available: http://mac.mellon.org/issues-in-conservation-documentation/Rembrandt%20Database%202-23-09.pdf [Accessed July 2014].

DREDGE, P. 2010. Collections of paint colour charts, paint tins and paintings as a source for developing an understanding of paint making history. *MUSEUMS AUSTRALIA NATIONAL CONFERENCE.* 45-50. Melbourne: http://www.ma2010.com.au/docs/ma2010_dredge.pdf.

DREDGE, P. 2012. A history of Australian housepaint technology from the 1920s to the 1950s, with reference to its use by Australian artists, particularly Sidney Nolan. *AICCM Bulletin,* 33**,** 53-61.

EIDE, Ø., FELICETTI, A., ORE, C., D'ANDREA, A. & HOLMEN, J. 2008. Encoding cultural heritage information for the semantic web. *In Procedures for Data Integration through CIDOC-CRM Mapping, EPOCH Conference on Open Digital Cultural Heritage Systems.* 1-7.

ELSAYED, I., MADEY, G. & BREZANY, P. 2011. Portals for collaborative research communities: two distinguished case studies. *Concurrency and Computation: Practice and Experience,* 23**,** 269-278.

EMBAREK, M. & FERRET, O. 2008. Learning patterns for building resources about semantic relations in the medical domain. *6th Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco.*

FEIGENBAUM, L., TORRES, E. & YUNG, W. 2006. *Javascript: SPARQL Service/Query* [Online]. Available: http://www.thefigtrees.net/lee/sw/sparql.js [Accessed July 2014].

FLICK, E. W. 2005. *Paint & Ink Formulations Database* [Online]. William Andrew Publishing. Available: http://www.knovel.com/web/portal/browse/display?_EXT_KNOVEL_DISPLAY_bookid=1197 [Accessed July 2014].

GATE. 2012. *Appendix G: Part-of-Speech Tags used in the Hepple Tagger* [Online]. Available: http://gate.ac.uk/sale/tao/splitap7.html#x37-743000G [Accessed July 2014].

GETTY. 2010a. *The Getty Conservation Institute* [Online]. Getty Research Institute, J. Paul Getty Trust. Available: http://www.getty.edu/conservation/ [Accessed July 2014].

GETTY. 2010b. *Getty Vocabularies* [Online]. Getty Research Institute, J. Paul Getty Trust. Available: http://www.getty.edu/research/tools/vocabularies/index.html [Accessed July 2014].

GOBLE, C., CORCHO, O., ALPER, P. & DE ROURE, D. 2006. e-Science and the Semantic Web: a symbiotic relationship. *Discovery Science.* 1-12. Springer.

GOBLE, C. A. & DE ROURE, D. C. 2007. myExperiment: social networking for workflow-using e-scientists. *Proceedings of the 2nd workshop on Workflows in support of large-scale science.* 1-2. ACM.

GOODALL, S., LEWIS, P., MARTINEZ, K., SINCLAIR, P., GIORGINI, F., ADDIS, M., BONIFACE, M., LAHANIER, C. & STEVENSON, J. 2004. SCULPTEUR: Multimedia Retrieval for Museums. *Image and Video Retrieval, 638-646. Springer Berlin Heidelberg.*

GRAY, J., LIU, D. T., NIETO-SANTISTEBAN, M., SZALAY, A., DEWITT, D. J. & HEBER, G. 2005. Scientific data management in the coming decade. *ACM SIGMOD Record,* 34**,** 34-41.

GREEN, D. & MUSTALISH, R. 2009. Digital Technologies and the Management of Conservation Documentation. A Survey Commissioned by the Andrew W. Mellon Foundation, New York.

GROUIN, C., ABACHA, A. B., BERNHARD, D., CARTONI, B., DELGER, L., GRAU, B., LIGOZAT, A. L., MINARD, A. L., ROSSET, S. & ZWEIGENBAUM, P. 2010. CARAMBA: concept, assertion, and relation annotation using machine-learning based approaches. *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2.*

GRUBER, T. R. 1995. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.,* 43**,** 907-928.

HASLHOFER, B., MOMENI, E., GAY, M. & SIMON, R. 2010. Augmenting Europeana content with linked data resources. *Proceedings of the 6th International Conference on Semantic Systems.* Graz, Austria: 1-3. ACM.

HERNÁNDEZ, F., RODRIGO, L., CONTRERAS, J. & CARBONE, F. 2008. Building a Cultural Heritage Ontology for Cantabria. *Annual Conference of CIDOC,* http://network.icom.museum/fileadmin/user_upload/minisites/cidoc/ConferencePapers/2008/64_papers.pdf.

HEYDENREICH, G. 2009. *The Cranach Digital Archive* [Online]. Andrew W. Mellon Foundation. Available: http://mac.mellon.org/issues-in-conservation-documentation/Cranach%20project%20description%20Nov%202009.pdf [Accessed July 2014].

HOBBS, J. R. & PAN, F. 2006. Time ontology in OWL. *W3C working draft,* 27**,** 133. http://www.w3.org/TR/owl-time/.

HOFMANN, M. 2009. *The National Gallery's Mellon Digital Documentation Project: The Raphael Research Resource* [Online]. Andrew W. Mellon Foundation. Available: http://mac.mellon.org/NGL%20Raphael%20Project%20update%203-3-09.pdf [Accessed July 2014].

HOHMANN, G. & SCHIEMANN, B. 2013. An ontology-based communication system for cultural heritage: Approach and progress of the WissKI project. *Scientific Computing and Cultural Heritage, 127-135. Springer Berlin Heidelberg.*

HOVY, E., KOZAREVA, Z. & RILOFF, E. 2009. Toward completeness in concept extraction and classification. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: 2 (2), 948-957.* Singapore: Association for Computational Linguistics.

HULL, D., WOLSTENCROFT, K., STEVENS, R., GOBLE, C., POCOCK, M. R., LI, P. & OINN, T. 2006. Taverna: a tool for building and running workflows of services. *Nucleic acids research,* 34 (suppl 2)**,** W729-W732.

HUNT, W. H. 2006. Materials informatics: Growing from the Bio World. *JOM,* 58**,** 88-88.

HUNTER, J. & ODAT, S. 2011. Building a Semantic Knowledge-base for Painting Conservators. *E-Science (e-Science), 2011 IEEE 7th International Conference on, 173-180.* IEEE.

HYVÖNEN, E. 2009. Semantic Portals for Cultural Heritage. *In Handbook on ontologies, 757-778. Springer Berlin Heidelberg*.

HYVÖNEN, E., MÄKELÄ, E., KAUPPINEN, T., ALM, O., KURKI, J., RUOTSALO, T., SEPPÄLÄ, K., TAKALA, J., PUPUTTI, K. & KUITTINEN, H. 2009. CultureSampo: A National Publication System of Cultural Heritage on the Semantic Web 2.0. *The Semantic Web: Research and Applications,* 5554**,** 851-856.

HYVÖNEN, E., MÄKELÄ, E., SALMINEN, M., VALO, A., VILJANEN, K., SAARELA, S., JUNNILA, M. & KETTULA, S. 2005. MuseumFinland—Finnish museums on the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web,* 3 (2)**,** 224-241.

HYVÖNEN, E., RUOTSALO, T., HÄGGSTRÖM, T., SALMINEN, M., JUNNILA, M., VIRKKILÄ, M., HAARAMO, M., MÄKELÄ, E., KAUPPINEN, T. & VILJANEN, K. 2006. Culturesampo–finnish culture on the semantic web: The vision and first results. *In Developments in Artificial Intelligence and the Semantic Web-Proceedings of the 12th Finnish AI Conference STeP, 26-27.* Berlin: LIT Verlag.

IBM. 2013. *Content Analytics - IBM LanguageWare* [Online]. Available: http://www-01.ibm.com/software/globalization/topics/languageware/ [Accessed July 2014].

INCCA. 1999. *INCCA Database for Artists' Archives (IDAA)* [Online]. Netherlands Institute for Cultural Heritage. Available: http://www.inccamembers.org/OCMT/ [Accessed July 2014].

IOANNIDES, M., ARNOLD, D., NICCOLUCCI, F. & MANIA, K. 2006. Extraction and mapping of CIDOC-CRM encodings from texts and other digital formats. *The 7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage.* Nicosia, Cyprus.

IRUG. 2010. *IRUG Spectral Database Edition 2000* [Online]. Infrared and Raman Users Group. Available: http://www.irug.org/search-spectral-database [Accessed July 2014].

JSTOR. 2000. *JSTOR Studies in Conservation* [Online]. Available: http://www.jstor.org/ [Accessed July 2014].

JURIJ, L., NATASA, M.-F., MARKO, G. & LESKOVEC, J. 2005. Extracting Summary Sentences Based on the Document Semantic Graph. Microsoft Research, Microsoft Corporation. http://www.cs.cmu.edu/~jure/pubs/nlpspo-msrtr05.pdf.

KIM, J. D., OHTA, T., TSURUOKA, Y., TATEISI, Y. & COLLIER, N. 2004. Introduction to the bio-entity recognition task at JNLPBA. *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications, 70-75.* Association for Computational Linguistics.

KOUTSOMITROPOULOS, D. A. & PAPATHEODOROU, T. S. 2007. Expressive Reasoning about Cultural Heritage Knowledge Using Web Ontologies. *Proc. of 3d International Conference on Web Information Systems and Technologies.* WEBIST (2), 276-281.

KRAFFT, D. B., CAPPADONA, N. A., CARUSO, B., CORSON-RIKERT, J., DEVARE, M. & LOWE, B. J. 2010. Vivo: Enabling national networking of scientists. In: Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, 1310-1313. Raleigh, NC, US.

KRALLINGER, M., LEITNER, F., RABAL, O., VAZQUEZ, M., OYARZABAL, J. & VALENCIA, A. 2013. Overview of the chemical compound and drug name recognition (CHEMDNER) task. *BioCreative Challenge Evaluation Workshop (2).*

KUDO, T., YAMAMOTO, K. & MATSUMOTO, Y. 2004. Applying conditional random fields to Japanese morphological analysis. *Proceedings of EMNLP, 4, 230-237.*

KURTZ, D., PARKER, G., SHOTTON, D., KLYNE, G., SCHROFF, F., ZISSERMAN, A. & WILKS, Y. 2009. Claros-bringing classical art to a global public. *e-Science, 2009. e-Science'09. Fifth IEEE International Conference on, 20-27.* IEEE.

LAGOZE, C. 2009. The oreChem project: Integrating chemistry scholarship with the Semantic Web. *Proceedings of the WebSci'09: Society On-Line, 18-20 March 2009.* Athens, Greece.

LAGOZE, C., VAN DE SOMPEL, H., NELSON, M. L., WARNER, S., SANDERSON, R. & JOHNSTON, P. 2008. Object Re-Use & Exchange: A Resource-Centric Approach. *ArXiv e-prints arXiv:0804.2273.*

LEONA, M. & DUYNE, R. V. 2009. *Chemistry and Materials Research at the Interface between Science and Art* [Online]. Report of a Workshop Cosponsored by the NSF and the Andrew W. Mellon Foundation, Arlington, Virginia. Available: http://mac.mellon.org/NSF-MellonWorkshop/2009%20NSF%20final%20report%20full%20res.pdf [Accessed July 2014].

LIAO, Z. & ZHANG, Z. 2012. A Generic Classifier-Ensemble Approach for Biomedical Named Entity Recognition. *In Advances in Knowledge Discovery and Data Mining,* 7301**,** 86-97. Springer Berlin, Heidelberg.

LIEBERMAN, J., SINGH, R. & GOAD, C. 2007. W3c geospatial vocabulary. *W3C Incubator Group Report,* 1-13. http://ontogenealogy.com/documents/2012/08/w3c-geospatial-vocabulary-2007.pdf.

LIN, C. Y. & HOVY, E. 2000. The automated acquisition of topic signatures for text summarization. *Proceedings of the 18th conference on Computational linguistics.* 1, 495-501. Association for Computational Linguistics.

LUDÄSCHER, B., ALTINTAS, I., BERKLEY, C., HIGGINS, D., JAEGER, E., JONES, M., LEE, E. A., TAO, J. & ZHAO, Y. 2006. Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience,* 18**,** 1039-1065.

MARTINEZ, K. & ISAKSEN, L. 2010. The semantic web approach to increasing access to cultural heritage. *In: Bailey, C. And Gardiner, H. (eds) Revisualizing Visual Culture. Farnham; Burlington VT: Ashgate.***,** 29-44. http://eprints.soton.ac.uk/268557/.

MAYNARD, D., FUNK, A. & PETERS, W. 2009. SPRAT: a tool for automatic semantic patternbased ontology population. In International conference for digital libraries and the semantic web, Trento, Italy.

MCCALLUM, A. & LI, W. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, 4.* Association for Computational Linguistics.

MCCALLUM, A. K. 2002. *MALLET: A Machine Learning for Language Toolkit* [Online]. Available: http://mallet.cs.umass.edu/ [Accessed July 2014].

MCGUINNESS, D. & HARMELEN, F. 2004. *OWL Web Ontology Language Overview* [Online]. W3C. Available: http://www.w3.org/TR/owl-features/ [Accessed July 2014].

MCI. 2012. *Smithsonian's Museum Conservation Institute* [Online]. Available: http://www.si.edu/mci/index.html [Accessed July 2014].

MCPHILLIPS, T., BOWERS, S. & LUDÄSCHER, B. 2006. Collection-oriented scientific workflows for integrating and analyzing biological data. *Data Integration in the Life Sciences.* 248-263. Springer.

MELLON. 2005. *Mellon Pilot Project: Open-source Solutions for Sharing Data and Images* [Online]. The Museum of Fine Arts, Boston. Available: http://mac.mellon.org/SH%20abstract%203-24-09.pdf [Accessed July 2014].

MELLON. 2007. *Mellon Pilot Project on Conservation and Science Documentation* [Online]. The British Museum. Available: http://mac.mellon.org/issues-in-conservation-documentation/the_british_museum_pilot.pdf [Accessed July 2014].

MELLON. 2009. *ConservationSpace* [Online]. Office of Digital Assets and Infrastructure, Yale University. Available: http://www.conservationspace.org/Home.html [Accessed July 2014].

MEURS, M. J., MURPHY, C., MORGENSTERN, I., NADERI, N., BUTLER, G., POWLOWSKI, J., TSANG, A. & WITTE, R. 2011. Semantic text mining for lignocellulose research. *Proceedings of the ACM fifth international workshop on Data and text mining in biomedical informatics, 35-42.* Glasgow, Scotland, UK: ACM.

MFA-BOSTON. 1997. *CAMEO Materials Database Search* [Online]. Museum of Fine Arts, Boston. Available: http://cameo.mfa.org/wiki/Category:Materials_database [Accessed July 2014].

MFA-BOSTON. 2010. *Forbes Pigment Database* [Online]. Museum of Fine Arts, Boston. Available: http://cameo.mfa.org/wiki/Forbes_Pigment_Database [Accessed July 2014].

MINTZ, M., BILLS, S., SNOW, R. & JURAFSKY, D. 2009. Distant supervision for relation extraction without labeled data. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2.* Suntec, Singapore: Association for Computational Linguistics.

MONACHESI, P., MARKUS, T. & MOSSEL, E. 2009. Ontology enrichment with social tags for elearning. *Learning in the Synergy of Multiple Disciplines,* 5794**,** 385-390.

MONICO, L., VAN DER SNICKT, G., JANSSENS, K., DE NOLF, W., MILIANI, C., DIK, J., RADEPONT, M., HENDRIKS, E., GELDOF, M. & COTTE, M. 2011. Degradation Process of Lead Chromate in Paintings by Vincent van Gogh Studied by Means of Synchrotron X-ray Spectromicroscopy and Related Methods. 2. Original Paint Layer Samples. *Analytical Chemistry,* 83**,** 1224-1231.

MONROY, C., FURUTA, R. & CASTRO, F. 2010. Using an ontology and a multilingual glossary for enhancing the nautical archaeology digital library. *Proceedings of the 10th annual joint conference on Digital libraries, 259-262.* Gold Coast, Queensland, Australia: ACM.

MOTIK, B., GRAU, B. C., HORROCKS, I., WU, Z., FOKOUE, A. & LUTZ, C. 2012. *OWL 2 Web Ontology Language Profiles (Second Edition)* [Online]. W3C. Available: http://www.w3.org/TR/owl2-profiles/#OWL_2_RL [Accessed July 2014].

MUNPEX. 2012. *Multi-lingual Noun Phrase Extractor (MuNPEx)* [Online]. semanticsoftware.info. Available: http://www.semanticsoftware.info/munpex [Accessed July 2014].

MYERS, D. 2010. *Color of Art Pigment Database* [Online]. Available: http://www.artiscreation.com/Color_index_names.html [Accessed July 2014].

NA, L., LEILEI, Z., PRASENJIT, M., KARL, M., ERIC, P. & GILES, C. L. 2010. oreChem ChemXSeer: a semantic digital library for chemistry. *Proceedings of the 10th annual joint conference on Digital libraries: ACM.* Gold Coast, Queensland, Australia: 245-254. ACM.

NAVIGLI, R. & VELARDI, P. 2008. From Glossaries to Ontologies: Extracting Semantic Structure from Textual Definitions. *Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge.* 71-87. IOS Press.

NEVIN, A. 2009. *The Master of the Fogg Pietá ~ Maestro di Figline Project* [Online]. Andrew W. Mellon Foundation. Available: http://mac.mellon.org/Master%20of%20the%20Fogg%20Pieta%20Abstract.pdf [Accessed July 2014].

NIST. 2011. *NIST Chemistry WebBook* [Online]. National Institute of Standards and Technology. Available: http://webbook.nist.gov/ [Accessed July 2014].

NIST. 2012. *Standard Reference Data* [Online]. Available: http://www.nist.gov/srd/ [Accessed July 2014].

OERC. 2014. *CLAROS* [Online]. Available: http://www.clarosnet.org/XDB/ASP/clarosHome/ [Accessed July 2014].

OINN, T., ADDIS, M., FERRIS, J., MARVIN, D., SENGER, M., GREENWOOD, M., CARVER, T., GLOVER, K., POCOCK, M. R. & WIPAT, A. 2004. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics,* 20**,** 3045-3054.

OLDMAN, D. 2010. *ResearchSpace: Content Management Systems – Evaluation Considerations* [Online]. The British Museum. Available: http://www.researchspace.org/file-cabinet/ContentManagementSystemsEvaluationNotes.pdf?attredirects=0&d=1 [Accessed July 2014].

OLDMAN, D., NORTON, B. & HAMMA, K. 2014. *ResearchSpace - a Digital Wunderkammer* [Online]. Available: http://www.researchspace.org/ [Accessed July 2014].

OSMOND, G. 2012. Zinc white: a review of zinc oxide pigment properties and implications for stability in oil-based paintings. *AICCM Bulletin,* 33**,** 20-29.

OSMOND, G., BOON, J. J., PUSKAR, L. & DRENNAN, J. 2012. Metal Stearate Distributions in Modern Artists' Oil Paints: Surface and Cross-Sectional Investigation of Reference Paint Films Using Conventional and Synchrotron Infrared Microspectroscopy. *Applied Spectroscopy,* 66**,** 1136-1144.

OSMOND, G., KEUNE, K. & BOON, J. 2005. A study of zinc soap aggregates in a late 19th century painting by R.G. Rivers at the Queensland Art. *AICCM Bulletin,* 29**,** 37-46.

PANDIT, S. & HONAVAR, V. 2010. Ontology-guided Extraction of Complex Nested Relationships. *Proceedings of the 2010 22nd IEEE International Conference on Tools with Artificial Intelligence.* 2, 173-178. IEEE Computer Society.

PCSS. 2014. *Poznan Supercomputing and Networking Center* [Online]. Available: http://www.man.poznan.pl/online/en/ [Accessed July 2014].

PESTIAN, J. P., BREW, C., MATYKIEWICZ, P., HOVERMALE, D. J., JOHNSON, N., COHEN, K. B. & DUCH, W. 2007. A shared task involving multi-label classification of clinical free text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing.* Prague, Czech Republic: 97-104. Association for Computational Linguistics.

PIRRÓ, G., MASTROIANNI, C. & TALIA, D. 2010. A framework for distributed knowledge management: Design and implementation. *Future Generation Computer Systems,* 26**,** 38-49.

PROTÉGÉ. 1997. *Protégé* [Online]. Stanford Center for Biomedical Informatics Research. Available: http://protege.stanford.edu/ [Accessed July 2014].

QI, Y., SZUMMER, M. & MINKA, T. P. 2005. Diagram Structure Recognition by Bayesian Conditional Random Fields. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05).* 2, 191-196. IEEE Computer Society.

RASKIN, R. G. & PAN, M. J. 2005. Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). *Computers & Geosciences,* 31**,** 1119-1125.

REID, R. & EDWARDS, P. 2009. Ourspaces-a social semantic web environment for eScience. *In: AAAI Spring Symposium: Social Semantic Web: Where Web 2.0 Meets Web 3.0, 67-68.*

RESNIK, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007.* *http://arxiv.org/abs/cmp-lg/9511007*.

REYNOLDS, D., SHABAJEE, P. & CAYZER, S. 2004. Semantic information portals. *Proceedings of the 13th international World Wide Web conference on Alternate track papers and posters.* 290-291. New York, NY, USA: ACM.

ROBERTS, A., GAIZAUSKAS, R. & HEPPLE, M. 2008. Extracting clinical relationships from patient narratives. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing.* 10-18. Association for Computational Linguistics.

ROY, A., FOISTER, S. & RUDENSTINE, A. 2007. Conservation Documentation in Digital Form: A Continuing Dialogue about the Issues. *Studies in Conservation,* 52 (4)**,** 315-317.

RUDENSTINE, A. Z. & WHALEN, T. P. 2006. Conservation documentation in digital form: A dialogue about the issues. *Getty Conservation Institute Newsletter,* 21**,** 26-28.

SALTON, G. & BUCKLEY, C. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management,* 24**,** 513-523.

SANDERSON, R. & VAN DE SOMPEL, H. 2010. Making web annotations persistent over time. *Proceedings of the 10th annual joint conference on Digital libraries.* 1-10. Gold Coast, Queensland, Australia: ACM.

SCHMIDT, K., SCHMITZ, H.-C. & WOLPERS, M. 2011. Developing a Network of Cultural Heritage Objects Repositories for Educational Purposes. *Metadata and Semantic Research,* 240**,** 337-348.

SCHNEIDER, G., KALJURAND, K. & RINALDI, F. 2009. Detecting protein-protein interactions in biomedical texts using a parser and linguistic resources. *In Computational Linguistics and Intelligent Text Processing, 406-417.* Springer Berlin Heidelberg.

SHAH, A. R., SINGHAL, M., KLICKER, K. R., STEPHAN, E. G., WILEY, H. S. & WATERS, K. M. 2007. Enabling high-throughput data management for systems biology: the Bioinformatics Resource Manager. *Bioinformatics,* 23**,** 906-909.

SIDHU, A., DILLON, T. S., CHANG, E. & SIDHU, B. 2007. Protein ontology development using OWL. *In OWLED, 188, 63-80.*

SMITH, D., OWENS, A., RUSSELL, A., HARRIS, C. & WILSON, M. 2005. The evolving mSpace platform: leveraging the semantic web on the trail of the memex. *Proceedings of the sixteenth ACM conference on Hypertext and hypermedia.* 174-183. Salzburg, Austria: ACM.

SMITH, V. S., RYCROFT, S. D., BRAKE, I., SCOTT, B., BAKER, E., LIVERMORE, L., BLAGODEROV, V. & ROBERTS, D. 2011. Scratchpads 2.0: a Virtual Research Environment supporting scholarly collaboration, communication and data publication in biodiversity science. *ZooKeys***,** 53. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3234431/.

STANDEVEN, H. A. 2011. *House Paints, 1900-1960: History and Use*, Getty Publications.

SUGIMOTO, S., BAKER, T. & WEIBEL, S. 2002. Dublin Core: Process and Principles. *Digital Libraries: People, Knowledge, and Technology,* 2555**,** 25-35.

THEODORIDOU, M., TZITZIKAS, Y., DOERR, M., MARKETAKIS, Y. & MELESSANAKIS, V. 2010. Modeling and querying provenance by extending CIDOC CRM. *Distributed and Parallel Databases,* 27**,** 169-210.

TOLEDO, M., CAPRETZ, M. & ALLISON, D. 2009. Recovering Brazilian Indigenous Cultural Heritage Using New Information and Communication Technologies. *IEEE/WIC/ACM International Joint Conferences (WI-IAT '09) on Web Intelligence and Intelligent Agent Technologies.* 3, 199-202.

TUDHOPE, D., MAY, K., BINDING, C. & VLACHIDIS, A. 2011. Connecting archaeological data and grey literature via semantic cross search. *Internet Archaeology,* 30 (5). http://dx.doi.org/10.11141/ia.30.5.

VAHUR, S. 2009. *FT-IR Spectra of Binders and Colorants* [Online]. Institute of Chemistry, University of Tartu, Estonia. Available: http://www.ut.ee/katsekoda/IR_Spectra/ [Accessed July 2014].

VAN OSSENBRUGGEN, J., AMIN, A., HARDMAN, L., HILDEBRAND, M., ASSEM, M. V., OMELAYENKO, B., TORDAI, A., SCHREIBER, G., BOER, V. D. & WIELINGA, B. 2007. Searching and Annotating Virtual Heritage Collections with Semantic-Web Techniques. *Museums and the Web 2007: Proceedings.* Toronto: Archives & Museum Informatics.

VLACHIDIS, A. 2012. *Semantic Indexing via Knowledge Organization Systems: Applying the CIDOC-CRM to Archaeological Grey Literature.* Doctoral dissertation, University of Glamorgan.

VLACHIDIS, A., BINDING, C., MAY, K. & TUDHOPE, D. 2013. Automatic metadata generation in an archaeological digital library: Semantic annotation of grey literature. *Computational Linguistics,* 458**,** 187-202.

W3C. 2008. *SPARQL Query Language for RDF* [Online]. W3C Recommendation 15 January 2008. Available: http://www.w3.org/TR/rdf-sparql-query/ [Accessed July 2014].

W&N. 2009. *Winsor & Newton Archive of 19th Century Artists' materials* [Online]. Hamilton Kerr Institute, University of Cambridge. Available: http://www-hki.fitzmuseum.cam.ac.uk/archives/wn/database.php [Accessed July 2014].

WIELEMAKER, J., HILDEBRAND, M., OSSENBRUGGEN, J. & SCHREIBER, G. 2008. Thesaurus-Based Search in Large Heterogeneous Collections. *The Semantic Web - ISWC 2008,* 5318**,** 695-708.

WU, Z. & PALMER, M. 1994. Verbs semantics and lexical selection. *Proceedings of the 32nd annual meeting on Association for Computational Linguistics.* 133-138. Association for Computational Linguistics.

YAMASHITA, F., FENG, C., YOSHIDA, S., ITOH, T. & HASHIDA, M. 2011. Automated Information Extraction and Structure−Activity Relationship Analysis of Cytochrome P450 Substrates. *Journal of Chemical Information and Modeling,* 51**,** 378-385.

ZHANG, S. & ELHADAD, N. 2013. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of biomedical informatics,* 46 (6)**,** 1088-1098.