

**SEVENTH FRAMEWORK PROGRAMME  
CAPACITIES**



**Research Infrastructures  
INFRA-2009-1 Research Infrastructures**

**OpenAIREplus**

**Grant Agreement 283595**

**“2nd-Generation Open Access Infrastructure for Research in  
Europe  
OpenAIREplus”**



**Connecting Data and Publications through e-Infrastructures**

Deliverable Code: D3.2

## Document Description

### Project

Title:	OpenAIREplus, 2 <sup>nd</sup> Generation Open Access Infrastructure for Research in Europe
Start date:	1 <sup>st</sup> December 2011
Call/Instrument:	INFRA-2011-1.2.2
Grant Agreement:	<b>283595</b>

### Document

Deliverable number:	D3.2
Deliverable title:	Connecting Data and Publications through e-Infrastructures
Contractual Date of Delivery:	31 <sup>th</sup> of January, 2013
Actual Date of Delivery:	09 <sup>th</sup> of January, 2014 (final version 2)
Editor(s):	
Author(s):	Florian Gräf, Maarten Hoogerwerf, Mathias Lösch, Jochen Schirrwagen, Sarah Callaghan, Paolo Manghi, Katerina Iatropoulou, Dimitra Keramida, Najla Rettberg
Reviewer(s):	Florian Gräf, Maarten Hoogerwerf, Jo McEntyre, Jochen Schirrwagen
Participant(s):	Alessia Bardi, Natalia Manola, Mateusz Kobos, Jo McEntyre, Oliver Kilian
Workpackage:	WP3
Workpackage title:	Studies on practices and principles of OA
Workpackage leader:	UNIBI
Workpackage participants:	NKUA, CNR, UNIWARSAW, EMBL, KNAW-DANS, STFC
Distribution:	Public
Nature:	Report
Version/Revision:	v2.0
Draft/Final:	Final
Total number of pages: (including cover)	
File name:	D3.2_connecting_data_publications_through_e- infrastructurestructures_v2_final
Key words:	Enanced Publications, Data-Publication Linkage

## Disclaimer

This document contains description of the OpenAIREplus project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) rules so, prior to using its content please contact the consortium head for approval.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the OPENAIRE consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 27 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors. (<http://europa.eu.int/>)



OpenAIREplus is a project funded by the European Union

## Table of Contents

<b>Document Description</b> .....	<b>2</b>
<b>Disclaimer</b> .....	<b>3</b>
<b>Table of Contents</b> .....	<b>4</b>
<b>List of Figures</b> .....	<b>5</b>
<b>List of Tables</b> .....	<b>6</b>
<b>Summary</b> .....	<b>7</b>
<b>1 Introduction</b> .....	<b>8</b>
<b>2 Identified Challenges and Opportunities</b> .....	<b>10</b>
2.1 Versions of a resource, dynamic resources, publication types.....	10
2.2 Meta-Information to present .....	10
2.3 Modeling aspects .....	10
2.4 Origin of Resources and Relationships .....	12
<b>3 Prototyping Analysis</b> .....	<b>13</b>
3.1 Life Sciences Prototype.....	13
3.2 Social Sciences and Humanities Prototype.....	22
<b>4 Discussion &amp; Conclusion</b> .....	<b>29</b>
<b>5 Outlook</b> .....	<b>31</b>
5.1 OpenAIRE – as a Registry for Data Citations.....	31
5.2 EuropePMC.....	31
<b>6 References</b> .....	<b>32</b>

## List of Figures

Figure 1-1. Graph of linked entities of a contextualized research result .....	8
Figure 2-1. Contextual entities related to a research article.....	11
Figure 2-2. Contextual entities related to a research dataset.....	11
Figure 3-1. Representation of database references in the EBI Webservice output .....	15
Figure 3-2. Different semantic types of text-mined terms in the Webservice.....	17
Figure 3-3. Reference list of a publication from the EBI Webservice.....	18
Figure 3-4. Entity-Relationship diagram designed for the Life Sciences demonstrator .....	19
Figure 3-5. NARCIS portal – the Dutch research information system.....	23
Figure 3-6. Example of an interview fragment defined in XML. It contains a transcription and metadata and it can be listened online.....	25
Figure 3-7. Screenshot of the demonstrator showing the description of a publication and positions references to its context on the left, right and bottom.....	27
Figure 5-1. Schematic workflow to lookup data citations.....	31

## List of Tables

Table 3-1. Example of RDF statements to describe relations between a publication, researchers and a project. 26

## Summary

The document reports results of the design, development and dissemination of “Subject-specific Pilots for Enhanced Publications” (T3.1). Being part of WP3 “Studies on practices and principles of OA”, the outcome of the task is twofold:

- (i) Development of three prototype applications to showcase how interconnected research information is being managed in different disciplines.
- (ii) Experiences and insights gained on a (potentially discipline-independent) implementation of systems capable of managing such linked artefacts that will inform the future development of the OpenAIRE infrastructure and its portal.

This report is based on the paper “Linking Data and Publications: Toward a Cross-Disciplinary Approach” from the same authors. It was presented at the International Digital Curation Conference, Amsterdam 2013 and submitted to the International Journal of Digital Curation.

# 1 Introduction

OpenAIRE aims to build a support structure for Open Access in Europe. OpenAIRE provides through its portal access to metadata about publications resulting from EC-funded research projects. It expands its scope to the entire European Open Access content from a network of institutional and thematic repositories. Running the Open Access Pilot in the EC's Seventh Framework Programme it allows project coordinators to make their scientific publications visible at a central place.

OpenAIRE will provide services allowing researchers, funding organizations and third party services to discover scientific results from a rich information space graph. As publications are not the only results of research projects, OpenAIRE extends its scope towards new forms of publications and considers research artefacts such as datasets and interlinks them with related entities in the portal. By interconnecting all related research objects OpenAIRE can represent datasets that have been described and cited in a given publication.

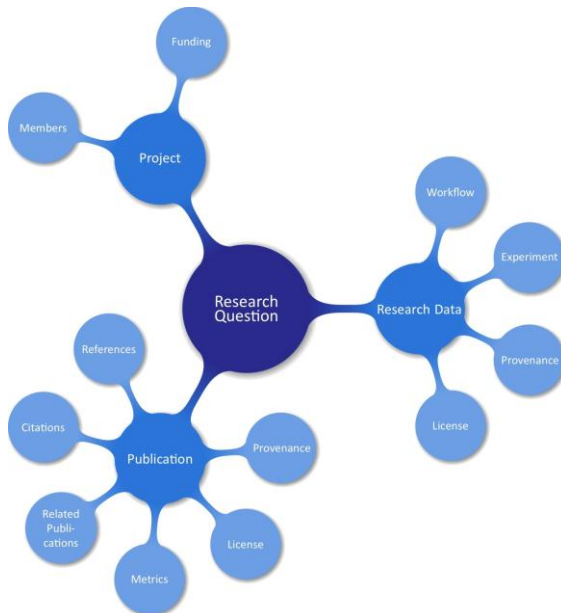


Figure 1-1. Graph of linked entities of a contextualized research result

The information space graph is built on the concept of Enhanced Publications. This concept has been explored in recent years by various projects and stakeholders. In the DRIVER-II project<sup>1</sup> "Enhanced Publications" were elaborated by demonstrators while providing an overview of a theoretical model and formulating a broad definition: "An Enhanced Publication is a publication that is enhanced with research data as evidence of the research, extra materials to illustrate or to clarify or post-publication data like commentaries and ranking" (Woutersen-Windhouver et al., 2009).

Open Access to publications and research data are vital requirements to put the concept of Enhanced Publications into practice. It is expected that this kind of publication will lead to improved visibility and impact of research data since they facilitate the re-use of scientific

<sup>1</sup> Digital Repository Infrastructure Vision for European Research: <http://cordis.europa.eu/projects/212147>



results by the combination of access to the data and their methodological description in the publication.

Two disciplines have been selected – the Life Sciences and the Social Sciences and Humanities. The current practices of researchers in these disciplines and their requirements on Open Access infrastructures are pointed out and described in the “Studies on Subject-Specific Requirements for Open Access Infrastructure” (Hogenaar et.al. 2011), (McEntyre 2011).

Though OpenAIRE aggregates metadata about publications and datasets from diverse disciplines, it must acknowledge that e.g. research practices, research data and the way that these relate to publications are very different across the disciplines. This faces OpenAIRE with the challenge to manage Enhanced Publications with different types of research objects and relations in a consistent way.

To learn from discipline-specific practices, OpenAIREplus works with three scientific partners that have a strong background in managing relations between research data and publications and data citation. These scientific partners are:

- The Dutch Data Archiving and Networked Services (DANS)
- The European Bioinformatics Institute (EMBL-EBI)
- The British Atmospheric Data Center (BADC) which is part of the Science & Technologies Facilities Council (STFC)

They contributed to the discussions on subject specific practices in managing publications and research data that further informed the design and development of demonstrators.

The demonstrators were aimed to serve for a better understanding of the diversity of research data used in selected disciplines. The focus is on:

- The kind of identifiers used to address research data
- How datasets are described in metadata
- The granularity of datasets and their citation
- Which APIs exist to query and fetch metadata about datasets in order to enrich existing bibliographic metadata with cross-links and context.
- How links between publications and datasets are managed
- If and how datasets are already cited in publications
- How the experiences gained from the demonstrators can be transferred into the overall development of the OpenAIRE infrastructure

This report describes the approach and outcomes of demonstrators using selected publications and research data of both disciplines.

## 2 Identified Challenges and Opportunities

The demonstrators serve multiple purposes: they should identify challenges, illustrate these, propose solutions and facilitate discussion within OpenAIRE, other e-infrastructures and research communities.

### 2.1 Versions of a resource, dynamic resources, publication types

Components of an Enhanced Publication may change over time. E.g. an author revises a publication or provides an erratum to a publication. Datasets may be updated as a result of data curation; more advanced algorithms for the computation of datasets might be available that could reduce error rates.

Resources that are referenced in an Enhanced Publication may originate in an online database and change dynamically without versioning information.

A research result can be published and disseminated in different ways, e.g. as an article published in a journal, as a poster at a conference or report in a workshop proceedings. Relations should be made on the different types of a publication, supported by proper controlled vocabulary terms.

### 2.2 Meta-Information to present

Comprehensive and qualitative metadata description is a decisive factor for the discovery of research results by aggregators and search engines.

In addition to basic bibliographic metadata (like creators, contributors, title, date, source, identifier) supplementary information becomes more important. Publications and associated research data are usually deposited in different repositories or databases. Therefore version information on the linked resource needs to be provided. Research data sets may be subject to privacy issues and provided as scientific use or public use files, depending on authenticated or anonymous users. Legal information on sharing and re-using of a specific resource of an Enhanced Publication is vital and have an effect on their processing, e.g. for text-mining.

### 2.3 Modeling aspects

Modeling an Enhanced Publication needs to consider different entities and their relations, which can be grouped by mandatory entities which can be enriched with other entities or linked for further exploration in subject-specific infrastructures, see Figure 2-1, Figure 2-2.

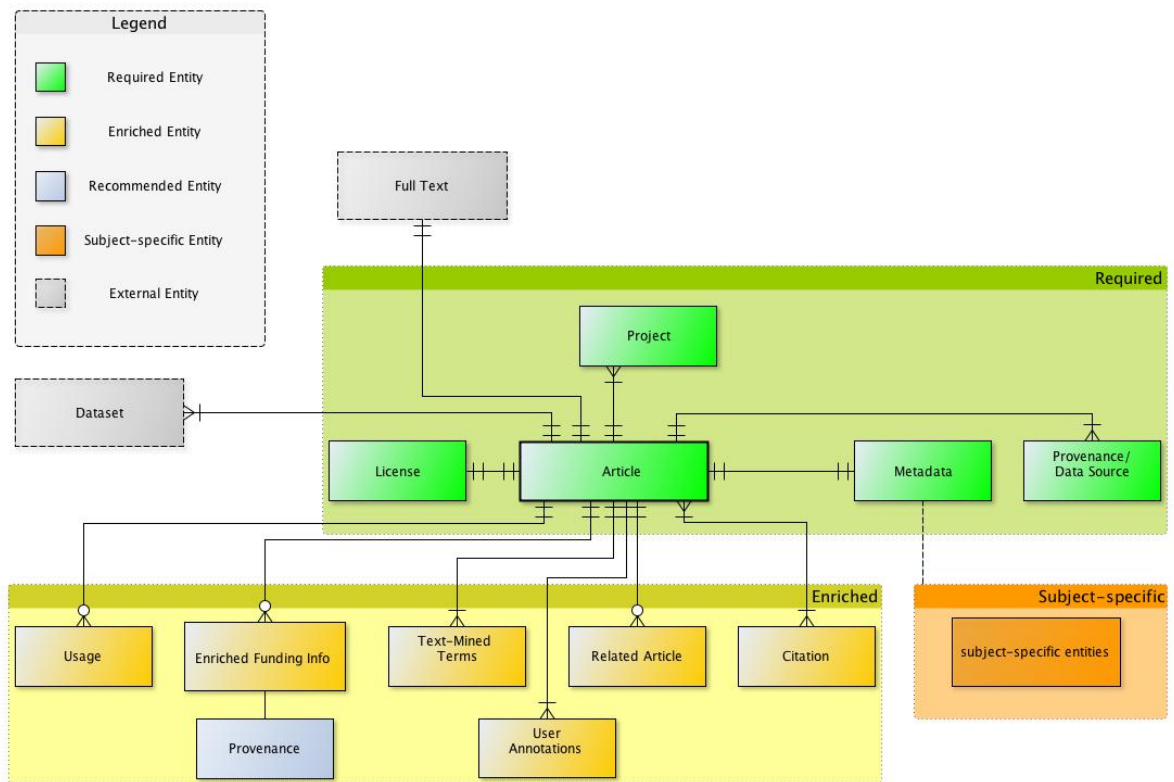


Figure 2-1. Contextual entities related to a research article

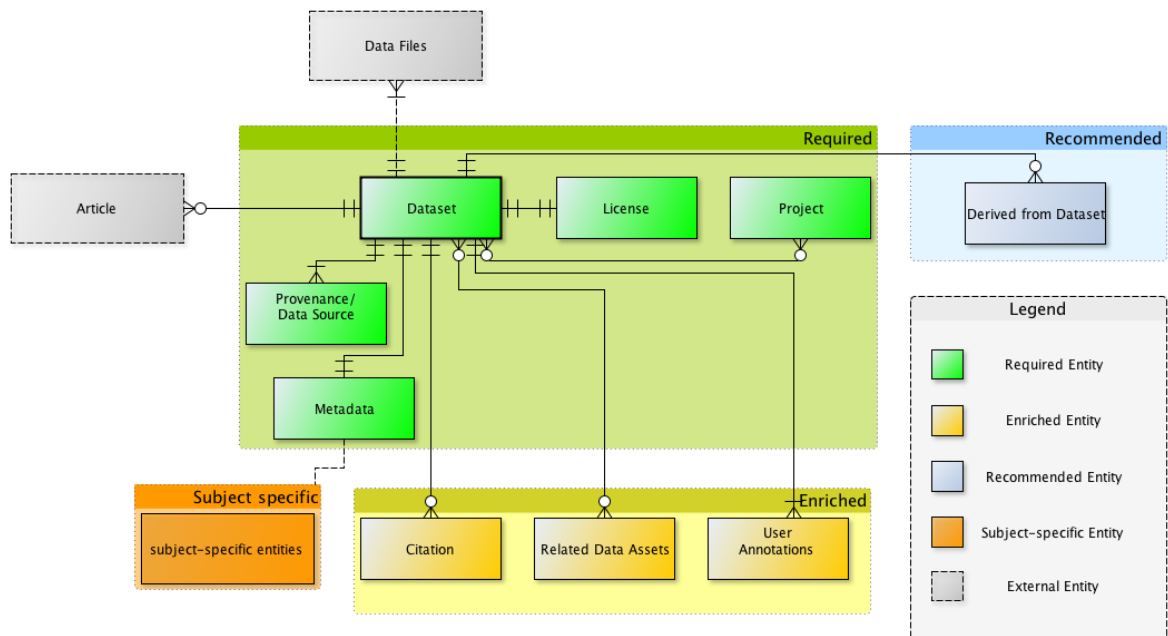


Figure 2-2. Contextual entities related to a research dataset

Other aspects affect so called nested aggregations. In this case an Enhanced Publication is an aggregate of another Enhanced Publication or parts of it.

Especially data set is a vaguely defined term, which can mean – in the context of “the long tail of science” - a container of a set of files or a data set of other data sets or e.g. a

segment of a dataset collection, e.g. a photo used in a paper in Archeology. A proper level of granularity needs to be defined to avoid information overload for the user.

## **2.4 Origin of Resources and Relationships**

Information about the origin of a resource and any modification should be recorded and made trackable.

It will give some indication of trustworthiness of the repository, database or archive.

## 3 Prototyping Analysis

This section presents two prototypes as an outcome of “Subject Specific Pilots for Enhanced Publications”.

### 3.1 Life Sciences Prototype

The landscape of life science methods and resulting data is characterized by diversity but at the same time by de facto standard experiment types and data structures allowing deposition in structured public databases. Consequently there are databases available for submission of data for most common experimental types. This is a typical way to support publications as data repositories provide a persistent identifier for the dataset that may be used to cite that dataset. This is a strongly applied method as it shows clear provenance of data used and generated in the course of the scientific work.

Even though life science studies produce data that may typically be assigned to a certain category and hence may be deposited into a corresponding repository, the data and repositories are diverse. The diversity of resource types, identifiers and the different ways of programmatic access to the data poses the main challenge of a project aiming to integrate data of manifold source and type.

The prototype uses Europe PubMed Central as source of publications in the Life Science domain and enriches the publication with FP7 project information retrieved using OpenAIRE, links of publications to each other (references and citations) and links to associated resources like the European Nucleotide Archive and Protein Databank in Europe to deposit sequencing results and Protein Structures.

The core idea of the Life Science Prototype<sup>1</sup> is to show how publications can be connected to related information that has already been collected and curated in a subject-specific infrastructure. Hence it will be elaborated in the following how data from EMBL databases was integrated with FP7 project information accessed using EuropePMC and OpenAIRE services.

#### 3.1.1 Europe PubMed Central

Europe PubMed Central is hosted by the EMBL-EBI (European Molecular Biology Laboratory and the European Bioinformatics Institute) and provides a variety of services and resources centered in the biomedical science domain.

As one of those Europe PubMed Central sits within this date context and services. It is based on the datasets of PubMed, PubMed Central and Agricola. PubMed, based on the journal citation database of the National Library of Medicine, provides more than 23 Million citations of life science journal publications, PubMed Central adds 2.8 Million full text articles and Agricola, the National Agricultural Library Catalog of the United States Department of Agriculture, supplies more than half a million agricultural citations.

Europe PubMed Central enriches publications from mentioned sources with links to data in the following ways:

---

<sup>1</sup> Life Sciences Demonstrator: <http://ub.unibi.de/oademo>

1. When a publication is cited by other publications then these citing publications are counted and displayed.
2. When a full text version of the publication is available it can be displayed in the EuropePMC Interface.
3. If a database links to a PubMed abstract, Europe PMC shows the reverse link i.e. it links from abstract to database.
4. Text mining is used to identify biological named entities and/or accession numbers in the text, which link back to related databases, as appropriate.

The connections obtained using these methods are not only visualized on the EuropePMC website but as well disseminated via Europe PMC's web services which were used to build the life science demonstrator.

### 3.1.2 Enhancing Publications

#### Identifying Suitable Publications

The very first step is to identify publications eligible for being related to Life Science datasets. Fortunately, the PubMed ID (PMID), as the identifier of MEDLINE based repositories, is a well-established identifier in this area. Therefore, only publications with associated PMIDs are imported in the prototype.

#### Querying the Europe PubMed Central Webservice

To enhance a life science publication it is reasonable to visualize the data which was obtained in the course a studies experiments. The EuropePMC web services allow the retrieval of such connections to other biological resources. For this purpose a PMID, once it is known can be used to query the Europe PubMed Central web service for related information. Currently, five types of information are obtained from EBI:

- **Database links:** references to biological database records holding links to publications
- **Text-mined terms:** biological named entities and accession numbers mined from the full text by EuropePMC that link to various resources at EBI/Europe PMC
- **References:** the publication's reference list
- **Citations:** citations from the EuropePMC citation database that cite a given publication
- **Full text links:** links to various instances of the publication with information on full text accessibility
- **MeSH headings:** Medical subject headings

#### Database Links

Database links are references to entries in EBI-hosted biological databases returned by the EBI web service for a given PMID. The reference is established via a so-called accession number, a unique ID in the scope of a database. While the accession number per se is not resolvable, it can be converted to a URI by the client through concatenation with its correct URL base path of the target database website. A database link record in the output of the

web service is furthermore augmented with some additional meta information, e.g., a human-readable description for display and further fields that are specific to the database. Figure 3-1 shows an example response for a query for database links with references to the EMBL and UniProt databases.

List of EBI databases linked to from the prototype:

- [ENA](#) formerly known as EMBL databank, the European Nucleotide Archive
- [UNIPROT](#) ( protein sequences)
- [INTACT](#) (molecular interactions)
- [CHEBI](#) (Chemical Entities of Biological Interest)
- [CHEMBL](#) (bioactive drug-like small molecules)
- [PDBe](#) (3D protein structures)
- [INTERPRO](#) (protein sequences and proteins sequence domains)

```
<?xml version="1.0" ?>
<dbCrossReferenceList>
  <dbCrossReference>
    <dbName>EMBL</dbName>
    <dbCount>10</dbCount>
    <dbCrossReferenceInfo>
      <info1>AF494760</info1>
      <info2>
        Arabidopsis thaliana ecotype Aa-0 actin-related
        protein 6 gene, partial cds.
      </info2>
      <info3>333</info3>
    </dbCrossReferenceInfo>
    ...
  </dbCrossReference>
  <dbCrossReference>
    <dbName>UNIPROT</dbName>
    <dbCount>63</dbCount>
    <dbCrossReferenceInfo>
      <info1>Q8GZK8</info1>
      <info2>Actin 8</info2>
      <info3>Arabidopsis thaliana</info3>
      <info4>UniProt</info4>
    </dbCrossReferenceInfo>
    ...
  </dbCrossReference>
</dbCrossReferenceList>
```

Figure 3-1. Representation of database references in the EBI Webservice output

EuropePMC gets the database links directly from the database records of the EBI hosted resources and uses them on the EuropePMC website to allow quick access of contextual resources.

### Text-mined Terms

Text-mined terms are references to biological databases automatically extracted from full texts by EBI via text mining. The identified terms can belong to one of the following domains:

- **accession** unique identifier given to database record
- **chemical** chemical name

- **disease** disease name
- **gene\_protein** protein identifier
- **go\_term** Gene Ontology (GO) term
- **organism** name of an organism

Depending on a term's domain (semantic type), links to different databases are established. For instance, chemical terms are converted to references to the ChEBI database. As with database links, the web-service does not return URIs but dataset IDs, so that the actual link targets have to be created by the client. Listed below are the link targets for the different semantic types of terms.

- **chemical:** *Query the CHEBI database:*  
`http://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:<accession>`
- **disease:** *Search in Europe PMC:*  
`http://europepmc.org/search/?query=<term>`
- **gene\_protein:** *Search in UNIPROT:*  
`http://www.uniprot.org/uniprot/?query=name:"<term>"`
- **go\_term:** *Search in UNIPROT:*  
`http://www.uniprot.org/uniprot/?query=go:<term>&sort=score`
- **organism:** *Search in UNIPROT:*  
`http://www.uniprot.org/uniprot/?query=organism:<term>&sort=score`

The nature of references produced by text-mined terms, except mined accession numbers, differs from database links in that they can be error-prone due to ambiguity problems. Furthermore, the resulting relations are not always as well-defined, because some references are implemented as search queries in the target databases. For instance, for the term "honey bees", which is mapped to the NCBI Taxonomy ID 7460 (honey bees). Following the cross-link triggers a query in the UniProt database for proteins of the organism 'honey bee' which results in a ranked hit list that may change over time.

The mined accession numbers, in contrast to the other types of text mined terms, are, like database links, well defined identifiers referring to known biological databases. They differ from database links in their mode of generation. While the database links are generated from biological resource records which contain crosslinks to publications, accession numbers are mined from full text publications.



```

<?xml version="1.0" ?>
<semanticTypeList>
  <semanticType>
    <name>chemical</name>
    <total>5</total>
    <tmSummary>
      <term>agarose</term>
      <count>1</count>
      <altNameList/>
      <dbName>chebi</dbName>
      <dbIdList>
        <dbId>2511</dbId>
      </dbIdList>
    </tmSummary>
    ...
  <semanticType>
    <name>gene_protein</name>
    <total>16</total>
    <tmSummary>
      <term>Nok</term>
      <count>2</count>
      <altNameList/>
      <dbName>uniprot</dbName>
      <dbIdList/>
    </tmSummary>
    ...
  </semanticType>
  <semanticType>
    <name>go_term</name>
    <total>16</total>
    <tmSummary>
      <term>binding</term>
      <count>13</count>
      <altNameList/>
      <dbName>GO</dbName>
      <dbIdList>
        <dbId>0005488</dbId>
      </dbIdList>
    </tmSummary>
    ...
  </semanticType>
</semanticTypeList>

```

Figure 3-2. Different semantic types of text-mined terms in the Webservice

## References and Citations

References and citations are exposed by the Webservice as bibliographic records with varying levels of detail. In most cases, they feature a persistent identifier in the `id` field, which can be used to create a link to their representations in either PubMed or EuropePMC.

```
<?xml version="1.0" ?>
<referenceList>
  <reference>
    <id>8502563</id>
    <source>MED</source>
    <citationType>JOURNAL ARTICLE</citationType>
    <title>
      Definition of a new alpha satellite suprachromosomal
      family characterized by monomeric organization.
    </title>
    <authorString>
      Alexandrov IA, Medvedev LI, Mashkova TD,
      Kisselev LL, Romanova LY, Yurov YB.
    </authorString>
    <journalAbbreviation>Nucleic Acids Res.</journalAbbreviation>
    <issue>9</issue>
    <pubYear>1993</pubYear>
    <volume>21</volume>
    <ISSN>0305-1048</ISSN>
    <ESSN>1362-4962</ESSN>
    <pageInfo>2209-2215</pageInfo>
    <citedOrder>1</citedOrder>
    <match>Y</match>
  </reference>
</referenceList>
```

Figure 3-3. Reference list of a publication from the EBI Webservice.

### Full Text Links and MeSH Headings

The Webservice lists the links to full text versions of a publication known to Europe PMC. If such links can be found, they are displayed in the demonstrator. Furthermore, if a publication is categorized by Medical Subject Headings (MeSH), these headings are listed with the publication metadata and fetched by the demonstrator. The MeSH terms are currently used to create search links to the Europe PMC portal, but could be used for an OpenAIRE search in a productive implementation.

### Querying OpenAIRE

To associate FP7-funded publications with the correct projects, the OpenAIRE Solr index is queried by the prototype with the title of the publication. If a publication can be matched to a project, the project metadata is imported into the prototype.

### Storing Retrieved Entities

Once the information has been fetched from EBI and OpenAIRE, the Life Science prototype represents each of the information types in its relational data model and links them to the publication.

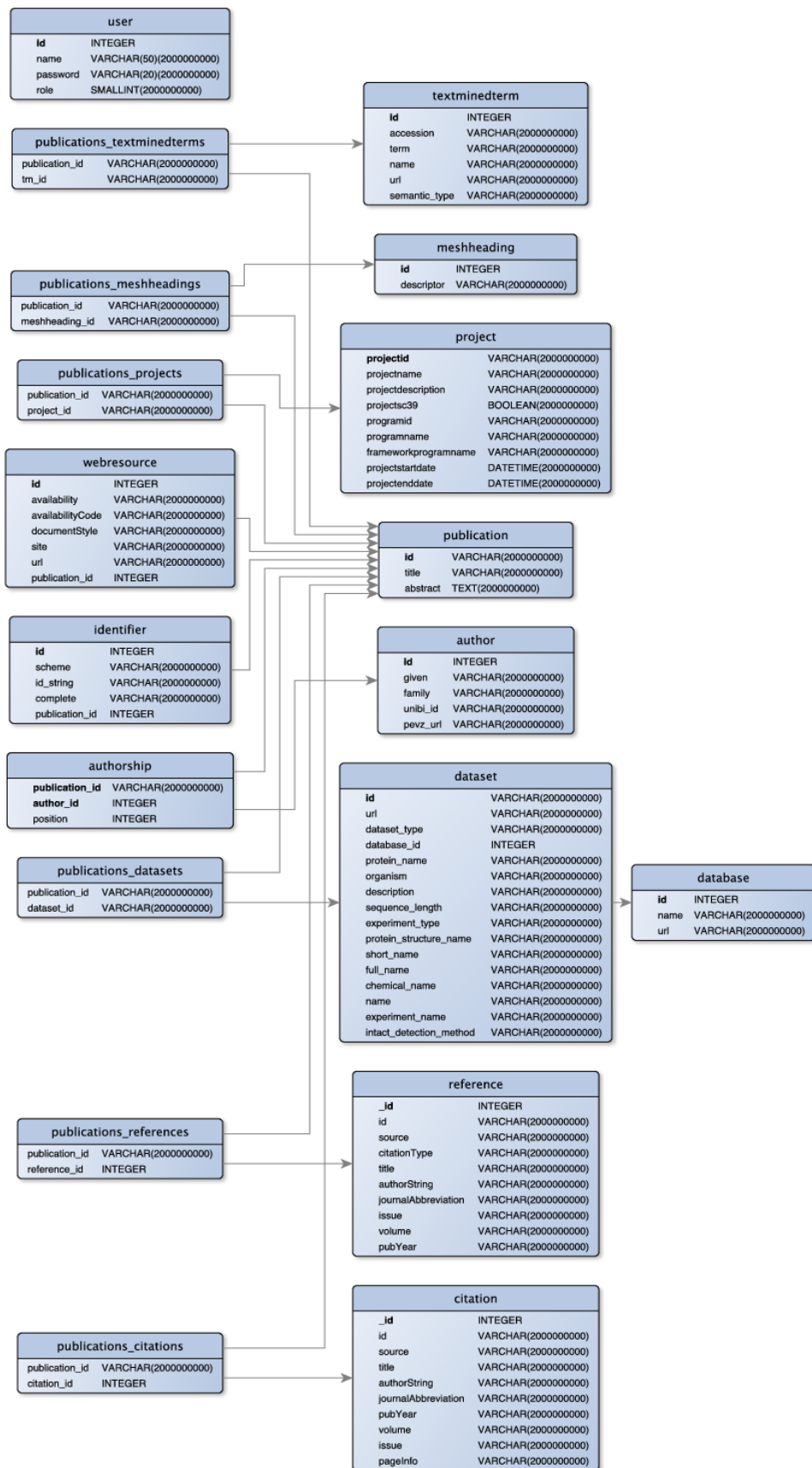


Figure 3-4. Entity-Relationship diagram designed for the Life Sciences demonstrator

### 3.1.3 The Demonstrator

#### Displaying Enhanced Publications

On the front end, the prototype features a simple web interface for displaying, browsing, and searching publications. It contains HTML splash pages for publications, projects, and datasets, with the following information:

##### Publications:

- Title
- Author(s)
- Abstract
- Database links (if available)
- Text-mined terms (if available)
- References (if available)
- Citations (if available)
- Project information (if available)

##### Datasets:

- ID
- Database
- Description
- Citing publications

##### Projects:

- Acronym
- Duration
- Resulting publications

#### Challenges and Opportunities

A particular challenge integrating OpenAIRE data with data from several life science sources was the use of domain specific API's to access data of interest. For this particular project the main API to aggregate data was the RESTful web service of Europe PubMed Central and had to be enriched by project information from OpenAIRE OAI-PMH based service. Even though the integration of interfaces to manifold external resources is a big challenge, EuropePMC, as a broker interfacing data of several heterogeneous databases via a single web service, simplifies the access.

To combine different types of data is not a challenge specific to this project. It is a rather general problem and especially the diversity of resources in the life science domain just augments this problem.



Opposing this challenge was the opportunity to gain experience integrating diverse life science data with project information from OpenAIRE into publication metadata. Especially expertise integrating data from EBI hosted resources was gained. Overall the demonstrator showcases how publication metadata enhanced by crosslinks to external resources could be presented in the OpenAIRE plus portal.

## 3.2 Social Sciences and Humanities Prototype

The landscapes for social sciences and humanities are different. The methods of quantitative social sciences show clear common methodologies whereas the qualitative social sciences and the humanities are very different. This is visible in the type of data that results from these disciplines. A large amount of data from the quantitative social sciences is numerical or survey data; the data from the qualitative social sciences and the humanities are far more heterogeneous: they use e.g. different sources from recorded interviews, paintings and books to digitized materials from e.g. an archaeological site. The study level is the most common data level: it is represented by a SPSS-like data file with documentation that represents e.g. a survey in the social sciences or it is represented by a set of various files that were collected during a project in the humanities. Persistent identification and citation of these datasets is becoming more common and usually happens at this study level. The need for more fine-grained identification (and citation) is also growing, but remains very challenging.

In the Netherlands, scientific publications are deposited in institutional repositories. The data can be shared via specific services and can be deposited for long-term preservation at institutions like DANS, Max Planck Institute for Psycholinguistics or Beeld & Geluid (Sound & Vision).

The social sciences and humanities prototype is built around NARCIS, a Dutch portal that aggregates content descriptions from various institutional repositories, CRIS systems and other sources. The basic idea of the prototype is to explore how interrelations between different types of scholarly resources can be captured and published. Two perspectives will be used: a narrower one following the concept of enhanced publications which are characterised by a defined aggregation of a publication with related resources, and a broader perspective of referenced resources that define each other's context. The prototype demonstrates how these can be visualized in a generic research portal.

### 3.2.1 NARCIS

NARCIS is a portal that provides access to scientific information from Dutch universities and other research organizations. The information includes descriptions about publications, datasets, research projects, researchers, organisations and (experimental) enhanced publications [see Figure 3-5].

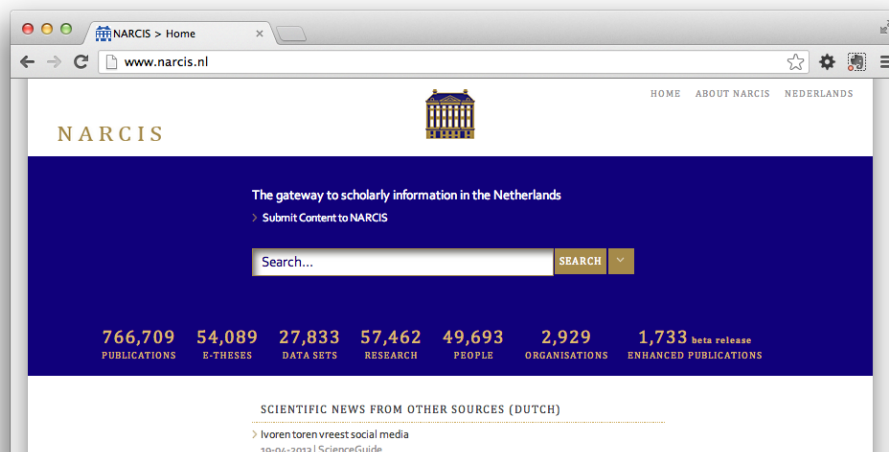


Figure 3-5. NARCIS portal – the Dutch research information system

DANS maintains the portal. It harvests the information on publications and datasets from Dutch institutional repositories and data archives (DANS, 3TU Datacentre, CentERdata<sup>1</sup> and Tilburg University). The information on research projects, people and organisations comes from the Dutch Research Databank (NOD), a database that is also maintained by editors from DANS. Some of the publications and datasets already contain identifiers<sup>2</sup> for their authors/creators. This allows these authors or creators to be connected to the researchers within the NOD, which also records researcher identifiers.

The enhanced publications in NARCIS are experimental results from tender projects from the Enhanced Publications programme by SURFshare<sup>3</sup>. These projects collaborated with researchers to enhance publications with different kinds of resources such as research data, project information, author information, visualizations, related publications etc. The enhancements are registered with the publications in the form of OAI-ORE resource maps (Lagoze et.al. 2008).

### 3.2.2 Enhancing Publications

For the pilot publications have been chosen that were already enhanced in two of the tender projects. The DataPlus project provides an example from the social sciences while the Veteran Tapes project provides an example from the humanities. The demonstrator shows these publications in a portal environment and adds the following references:

- **Survey elements:** references to DDI3 concepts and variables that were used for the publication (DataPlus publications only).
- **Interview fragments:** references to streaming audio fragments that are discussed in the publication (Veteran Tapes publication only).
- **Publications:** References to related publications (e.g. by citation) in NARCIS.

<sup>1</sup> <http://www.centerdata.nl/en/about-centerdata>

<sup>2</sup> Digital Author Identifier (DAI):

<http://www.surf.nl/en/themas/openonderzoek/infrastructuur/pages/digitalauthoridentifierdai.aspx>

<sup>3</sup> <http://www.surf.nl/en/themas/openonderzoek/verrijktepublicaties/Pages/default.aspx>

- **Datasets:** References to datasets registered in NARCIS that were used for this publication.
- **Researchers:** References to identified authors or editors of the publication in NARCIS.
- **Organizations:** References to identified organizations related to the publication in NARCIS.
- **Projects:** References to the funding projects in NARCIS.

The pilot also enhances objects that are related to publications. E.g. a referenced dataset is enhanced with other related publications.

## Survey Elements

The survey elements are DDI3-formatted concepts and variables that are used for the publication. The DataPlus project created a tool that allowed researchers to select their publication, browse a catalogue of studies, concepts and variables and select these. The result is stored and available as DDI3 formatted documents. The following is an example of a variable that is described in DDI3 format:

```
<l:Variable urn="urn:ddi:de.gesis:VariableScheme.ZA3811_VarSch.1.0.0:Variable.V33.1.0.0" id="ZA3811_V33">
  <r:Label xmlns:r="ddi:reusable:3_1">V33</r:Label>
  <r:Description xmlns:r="ddi:reusable:3_1">do you work unpaid
  for:tradeunions (Q5D)</r:Description>
</l:Variable>
```

These are connected to the publications using OAI-ORE Resource maps, which can be harvested using OAI-PMH from the repository. The OAI-PMH endpoint for the DataPlus project could no longer be reached. The DDI-information has been merged with the EVS portal. The metadata can be retrieved there:

HTML Example: <http://evs.uvt.nl/id/evs-uvt-nl:oai:evs.uvt.nl:3256420>  
OAI-ORE Example: <http://evs.uvt.nl/ore/evs-uvt-nl:oai:evs.uvt.nl:3256420>  
DDI3 Example: <http://evs.uvt.nl/ddi3/evs-uvt-nl:oai:evs.uvt.nl:3256420>

The demonstrator can, when available, load the DDI3-file and transform these into HTML using XSLT. As can be seen in the example, the variables and concepts are assigned unique DDI URN identifiers. This allows the demonstrator to locate instances of this variable within DDI3-enabled services.

## Interview Fragments

Researchers from different disciplines selected the interview fragments and referenced them from within their articles. These fragments were cut from the complete interviews as the latter may contain sensitive privacy information. This was done using a dedicated tool that allows users to select, describe and publish these fragments. The articles were published in a book. The PDF version of the book (van den Berg, 2010) contains references to the streaming fragments. An OAI-ORE resource map was created to connect these fragments to the metadata of the publication. An example of the fragments can be seen in Figure 3-6 . This example can be accessed here:

[http://www.watveteranenvertellen.nl/fragmenten/335\\_2133.57\\_2179.76.xml](http://www.watveteranenvertellen.nl/fragmenten/335_2133.57_2179.76.xml)



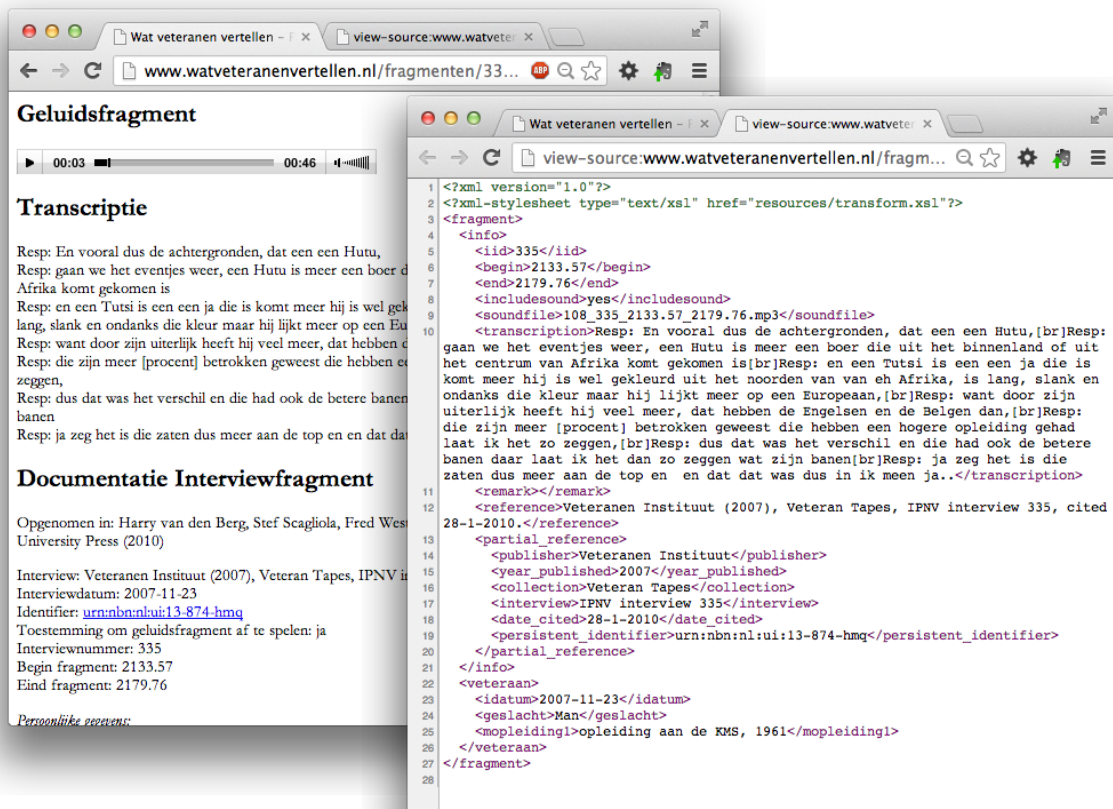


Figure 3-6. Example of an interview fragment defined in XML. It contains a transcription and metadata and it can be listened online.

The Resource Map was manually added to the demonstrator. It is loaded when the corresponding publication is displayed. XSLT is used to transform the resource map into HTML.

## Other Entities known by NARCIS

The issue with the above enhancements is that a portal like NARCIS cannot manage these: they are not supported by the repository infrastructure yet, resulting in the lack of standardized metadata, identification and access. Integrating them within a sustainable repository infrastructure requires these items to be managed by a data repository that can ensure long-term access to these fragments. Archives are upgrading their service levels but cannot yet support such fine-grained access to many different types of data.

What can be achieved? NARCIS currently contains descriptions of publications, datasets, researchers, projects and organizations. Unfortunately these are registered in isolated silos with hardly any (explicit) relations between them. There are no relations between publications, between publications and datasets or publications and projects. The only relations made are between researchers and publications or datasets via the Digital Author Identifiers (DAI). Many of the publication and dataset descriptions contain identified authors. In addition, many research profiles in NARCIS contain a DAI-number as well. This allows the dataset, publication and researcher descriptions to be interlinked by the

demonstrator. Note that NARCIS already supports this as well. Other references could not be automatically retrieved.

### Other Entities manually added to NARCIS

Other types of relations were retrieved by manually analysing the identified publications. This resulted in e.g. the distinction between the authors and the editor of a publication, the datasets of surveys or interviews and the related projects. Other publications were identified as similar publications, as these were deposited into other repositories, similar because these were by the same authors on the same topic or because they cited the publication.

The retrieved relations were registered as RDF triples. These triples used OAI-identifiers to identify publications and datasets, DAIs to identify researchers and internal numbers from NARCIS were used to reference projects and organizations. Note that many of the publications and datasets also contain persistent identifiers. The OAI-identifiers were used for practical reasons. To maintain a durable network of relations it will become important to register such relations with persistent identifiers. This ensures that the context is preserved when repositories are reorganised and it can also deal with redundant records in infrastructures.

Table 3-1. Example of RDF statements to describe relations between a publication, researchers and a project.

Subject	Predicate	Object
publication:oai:aup.nl:72413	doap:editor	dai:136423701
publication:oai:aup.nl:72413	doap:author	dai:102167834
publication:oai:aup.nl:72413	doap:fromProject	project:12345
publication:oai:aup.nl:72413	...	...

Retrieving these relations took a lot of time and the quality is sometimes poor because it was difficult to distinct different researchers or projects and because the citations often lack a persistent identifier.

Ideally this information is provided by the source itself. In many cases these are the researchers themselves. The author can identify its co-authors, the specific publications that they cited and the funders and organizations that facilitated the research. Repository staff could also provide this, though with less certainty. Unfortunately neither the researchers nor the librarians have time to provide this information. Further integration or optimization of repository systems and CRIS systems could minimize the required efforts: upon deposit of the publications, the system may already know the projects that the author has worked on. Citation information is more difficult but not unfeasible when e.g. unique identifiers are used.

Ultimately one could fetch these relations using data mining. These techniques can provide author disambiguation, retrieval of citations, etc. Unfortunately these techniques do not yet perform well enough: they only find parts of the relations and they find incorrect relations.

### 3.2.3 The Demonstrator

The demonstrator was constructed using a clone of NARCIS that could be extended with a mechanism to interrelate resources and a custom user interface that allows displaying and navigating these references.

PUBLICATION

WELFARE STATE EFFECTS ON SOCIAL CAPITAL AND INFORMAL... (2005) [Open access](#)


<p><b>Research</b></p> <hr/> <p><i>Input</i></p> <p>European Value Studies &gt;</p> <hr/> <p><b>Persons</b></p> <hr/> <p><i>Author</i></p> <p>Halman, Dr. L.C.J.M. &gt;</p> <p>Halman, L.C.J.M. &gt;</p> <p>Oorschot, W.J.H. van &gt;</p> <p>Arts, W.A. &gt;</p> <hr/> <p><b>Organisations</b></p> <hr/> <p><i>Repository</i></p> <p>Tilburg University &gt;</p>	<p>Title Welfare state effects on social capital and informal solidarity in the European Union: evidence from the 1999/2000 European Values Study</p> <p>Published in Policy and Politics: Studies of local government and its services, Vol. 33, No. 1, p.33-54. ISSN 03055736.</p> <p>Date 2005</p> <p>Type article</p> <p>Publication &gt; <a href="http://evs.uvt.nl/id/evs-uvt-nl:oai:evs.uvt.nl:3256420">http://evs.uvt.nl/id/evs-uvt-nl:oai:evs.uvt.nl:3256420</a></p> <p><u>Persistent Identifier</u> &gt; <a href="urn:nbn:nl:ui:12-3256420">urn:nbn:nl:ui:12-3256420</a></p> <p>Metadata &gt; XML</p> <p style="text-align: center;"> TILBURG UNIVERSITY</p> <p><b>External Database Links</b></p> <p><b>Concept: ethnocentrism</b>          Variable: V59: dont like as neighbours: muslims (Q7H)          Variable: V60: dont like as neighbours: immigrants/foreign workers (Q7I)          Variable: V64: dont like as neighbours: jews (Q7M)</p>	<p><b>Data</b></p> <hr/> <p><i>Cites</i></p> <p>European Values Study 199... &gt;</p> <p>EVS'99/2000: Release I... &gt;</p> <hr/> <p><b>Publications</b></p> <hr/> <p><i>Cites</i></p> <p>Welfare States, Solidarit... &gt;</p> <p>Three worlds of welfare c... &gt;</p> <p>Who should get what, and ... &gt;</p> <p>Individual motives for co... &gt;</p> <p>Wij en zij in Europa: De ... &gt;</p> <hr/> <p><i>Cited by</i></p> <p>Making the difference in ... &gt;</p> <p>Culture and social policy... &gt;</p> <p>Multi-level determinants ... &gt;</p>
--	---	---

Figure 3-7. Screenshot of the demonstrator showing the description of a publication and positions references to its context on the left, right and bottom.

The demonstrator publishes the requested publication in the centre of the page with all the references to the contextual resources on the left (projects, researchers and organisations) and on the right (datasets and publications). The specific enhancements are presented beneath the metadata of the publication. These enhancements link to the variables of fragments outside of the portal environments.

Beneath the fragments is a section that allows the description of the provenance. In NARCIS it is still feasible to describe on a general level where it gets its project information, or to provide a link to the source of the publication. This becomes more complex when the descriptions are no longer provided by one source alone: others can add contextual references to items that are provided by again other sources. In addition, these sources may update their description.

### 3.2.4 Lessons Learned and Feedback

Identifiers are key. They allow resources to be referenced. Publications and datasets are often provided with persistent identifiers on a global level. Several identification systems also exist for researchers but they are relatively new and not yet widely established. Currently multiple identifier schemes exist for researchers (e.g. DAI, ORCID, ResearcherID), but also for publications (e.g. URN, DOI, Handle). This makes de-duplication of the resources more complex: researchers having multiple identifiers, or publications getting a separate identifier in every repository. Universal schemes for projects or organizations are not known yet.

The 'by reference' model allows to define resources and relationships by help of authority information: funders provide project information, registries for repositories (OpenDOAR, re3data, databib) provide profile information about data sources, registration agencies for

persistent identifiers allow persistent identification of publications and datasets, identification systems for researchers allow them to create their own profiles. The drawback is that one is able to reference items only that are identifiable and have descriptions or profile information. For a portal like NARCIS or OpenAIRE this implies that one cannot easily reference resources that are outside the scope of such authorities. On the other hand: it does avoid conflicts of interest and a lot of moderation.

Displaying the discipline-specific metadata is not straightforward. It requires experts to define what information is relevant at all, and how this can be displayed in a useful way. Feedback told us that neither the listed fragments nor the listed variables were useful to them. Though the variables and concepts could help determine whether a specific publication fits the interest of somebody, the fragments could not support such judgement. The fragments are only relevant within the publication, where they can be evaluated within their context. An alternative purpose, that is not demonstrated, is to index these resources. This would allow researcher to refine their search-terms (faceted browsing) or to discover publications that use similar variables.

### **3.2.5 Challenges and opportunities**

Enhancing publications provides opportunities in terms of contextual discovery and interpretation. However, there is very little structure for these enhancements. The specific enhancements (variables and interview fragments) cannot yet be provided by generic infrastructure. DDI3 infrastructure for social science data is still in development. Video fragments are not yet supported by an archive like DANS, or on a standardized level by others.

If the scope is limited to more manageable resources like publications, datasets, projects, authors and organizations than there are still enough opportunities for discovery and interpretation. However, even on this level there remain many challenges. The identification of publications and datasets within repositories is becoming common practice, but there is little coordination: multiple instances get different identifiers and e.g. versioning is not yet dealt with. This leads to duplication and undocumented changes.

There is little motivation for providing the necessary relations, due to a lack of time from either the researchers themselves or the repository-staff. The workflows may be optimized to minimize the required effort by integrating CRIS systems, repositories and funding agencies. When the required efforts are minimized, proper policies may require researchers to provide with additional information that is missing.

The demonstrators did not yet model the relations, as the relations that could be retrieved were ad hoc. Existing models like CERIF will be applicable to many of the relations.

## 4 Discussion & Conclusion

The examples in the demonstrators share general concepts, such as publications, datasets, projects and persons (researchers). They interlink and allow users to navigate between them. The bibliographic metadata on all examples were fetched from their original sources without modifications. They could be harvested by OAI-PMH or queried by dedicated APIs. The linking of the resources depends on their identification. Commonly (but often subject-specific or regional limited) identifiers, such as PubMed IDs, DDI3-identifiers and DAIs, are good examples of well-defined identifiers that support automatic linking.

It turns out that more standardized application of globally unique and persistent identifiers for resources are needed, which includes a standardized, globally unique naming scheme for project information and researchers. Initiatives like ORCID<sup>12</sup> – introducing unique identifiers for researchers, DataCite<sup>3</sup> - introducing persistent identifiers and metadata kernel for research data sets, CrossRef<sup>4</sup> - introducing persistent identifiers for research publications tend in this direction.

It is important to note that the assignment of identifiers is taken care of by the appropriate stakeholder at the appropriate moment in the research or publication workflow. They allow others to define trusted and stable references to these resources.

An important difference between subject-specific examples can be observed in the interpretation of a dataset. Bio-Entities in the databases of Life Science infrastructures, as well as DDI3- encoded data in databases of the Social Sciences, are very well structured. They allow detailed identification and provide relations to other entities in their metadata. However, their internal structures differ from each other. The Humanities' data is the most heterogeneous, where the common structure is merely on the level of files and folders. Feedback from researchers and repository managers about the demonstrators indicate that access to detailed data entities is not of primary concern. It is more important that a researcher can discover other publications and data sources related to e.g. a concept in a questionnaire, rather than being able to analyse it from the portal of a generic infrastructure. This raises an important challenge: how can a cross-disciplinary infrastructure provide different subject-specific indexing for all its resources?

Feedback also recommends another valuable feature regarding the discovery of entities via categorizations in terms of detailed academic discipline, time, space and persons.

The relations among the different objects can be captured in different ways. It is recommended that they are captured early in the workflows by the most knowledgeable stakeholder, usually the author or creator of a resource. Of course it implies the use of well-established identifiers for the items to be referenced. The use of vocabularies and facilities for smart, auto-complete forms can support the establishment of those relations. To overcome the absence of relations among existing materials, methods for automatic

---

<sup>1</sup> ORCID: <https://orcid.org>

<sup>2</sup> About common and different aspects of ORCID and ISNI: [http://www.isni.org/isni\\_and\\_orcid](http://www.isni.org/isni_and_orcid)

<sup>3</sup> DataCite: <http://datacite.org>

<sup>4</sup> CrossRef: <http://www.crossref.org>

association of digital objects that have been explored recently should be investigated in the context of the OpenAIREplus project (Boland et al., 2012).

The process of constructing the demonstrators was only the first step towards a common approach of cross-disciplinary interlinking of research entities.

Further discussion and more examples (from other disciplines and infrastructures) are needed in order to acquire an at most complete picture of the richness and complexity of research outcome. The demonstrators support this discussion by providing concrete examples. They are a first but fundamental step towards integration of publication and research data services with other e-infrastructures to build the data continuum (Bird et.al. 2013).

## 5 Outlook

This section presents two scenarios which can take advantage of the future OpenAIREplus infrastructure.

### 5.1 OpenAIRE – as a Registry for Data Citations

Currently there is no established workflow for data repositories getting notified when a dataset stored with them gets cited in a publication. However, this information would be extremely valuable for creating back links for data citations (“cited-by”). As a solution, a central registry would need to be established that keeps track of datasets, publications, and their connections. Since OpenAIRE already attempts at creating such interlinked representations, it could fulfil this role, e.g., by exposing the stored connections through a web service. A possible workflow for this is illustrated in the following figure.

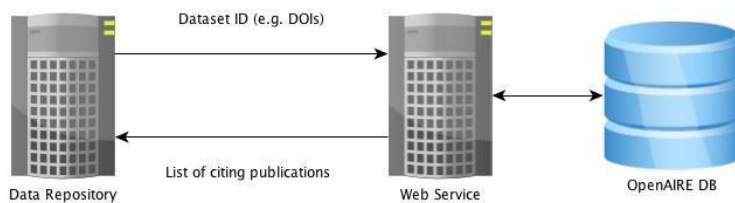


Figure 5-1. Schematic workflow to lookup data citations

A possible implementation scenario can be investigated in cooperation with the project “Peer REview for Publication & Accreditation of Research Data in the Earth sciences” (PREPARDE<sup>1</sup>).

### 5.2 EuropePMC

OpenAIRE aggregates from a large set of Open Access repositories of which some contain papers in the field of Life Sciences. Such papers are assigned with a PMID or PMCID.

OpenAIRE services allow the identification of such papers from its Information Space. Together with enriching of project information (if available) OpenAIRE can notify EuropePMC which in turn can link to the respective papers in the repositories.

Expectations are of improved visibility of the institutional repositories as well as of enriched contextual information about the research results which may increase use and re-use of publications linked to the bio-entities.

<sup>1</sup> <http://www2.le.ac.uk/projects/preparde>

## 6 References

- Bird I. et al.: A Vision for a European e-Infrastructure for the 21<sup>st</sup> Century. PrePrint. Retrieved from <http://cds.cern.ch/record/1550136> (2013) [accessed 13-June-2013]
- Berg H., Scagliola S.I., Wester F.P.J. (Eds.): Wat veteranen vertellen: Verschillende perspectieven op verhalen over ervaringen tijdens militaire operaties. Amsterdam University Press. Retrieved from <http://dx.doi.org/10.5117/9789085550341> (2010) [accessed 13-June-2013]
- Boland K., Ritze D., Eckert K., Mathiak B.: Identifying references to datasets in publications. In P. Zaphiris, G.Buchanan, E. Rasmussen, F. Loizides (Eds.), Lecture Notes in Computer Science: Vol. 7489. Theory and Practice of Digital Libraries (pp. 150-161). Springer (2012)
- Hogenaar A., Tjalsma H., Priddy M.: Research in the Humanities and Social Sciences. In: Meier zu Verl C, Horstmann W. (eds.): Studies on Subject-Specific Requirements for Open Access Infrastructure. Bielefeld: Universitätsbibliothek; 2011. Retrieved from <http://dx.doi.org/10.2390/PUB-2011-7> [accessed 13-June-2013]
- Lagoze C., Van de Sompel H., Nelson M., Warner S., Sanderson R., Johnston P.: Object reuse & exchange: A resource-centric approach. 2008. Retrieved from <http://arxiv.org/abs/0804.2273v1>
- McEntyre J., Ananiadou, S., Andrews, S., Black, W., Boulderstone, R., Buttery, P., Chaplin, D., Chevuru, S., Copley, N., Coleman, L. et al. (2011). UKPMC: A full text article resource for the life sciences. Nucleic acids research 39(suppl 1), D58–D65. <http://dx.doi.org/10.1093/nar/gkq1063>
- McEntyre J., Swan A.: Health Sciences. In: Meier zu Verl C, Horstmann W. (eds.): Studies on Subject-Specific Requirements for Open Access Infrastructure. Bielefeld: Universitätsbibliothek; 2011. Retrieved from <http://dx.doi.org/10.2390/PUB-2011-9> [accessed 13-June-2013]
- Woutersen-Windhouver S. et al.: Enhanced Publications: Linking Publications and Research Data in Digital Repositories. 2009. Retrieved from <http://dx.doi.org/10.5117/9789089641885> [accessed 13-June-2013]