Shout LOUD on a road trip to FAIRness: experience with integrating open research data at the Bibliotheca Hertziana

Alessandro Adamou

Introduction and background

I would like to thank Andrew [Hopkins], for the invitation and the great introduction. I would, in fact, also like to thank Tristan [Weddigen] and to all the other presenters before me for making my job easier, not having to explain so many things - although, yes, Andrew was right: there are many acronyms - you have seen already one or two - but there are still so many more I will need to explain. What I would like to talk about here is the job that is currently underway, and my experience with it, as we try and make the data in the various project of the Biblioteca Hertziana more compliant to data publishing principles, trying to reach out to a particular brand of Linked Open Data,[1] and to do so 'without disturbing the driver', meaning, without disrupting the editorial workflows of the project leads and collaborators, who are currently still working on live projects.

There are going to be some actual examples with code in various planes. There will be some RDF[2] code, some SPARQL[3] code, some XML[4], some programming code too. Why, I hear you say, would I be doing this? There is a funny quote that I always use on such occasions, taken from one of the Batman films by Christopher Nolan.[5] There's one scene where the character played by Morgan Freeman goes to Bruce Wayne and says, 'I have synthesised an antidote for this poison' and then he goes on giving a very complex explanation on what he went through to do it. When asked by Bruce Wayne, 'Am I supposed to understand any of that?', he replies 'No, I just wanted you to understand how hard it's been'. This is exactly not what I want to do here with you. The code is there exactly because I

[1] Tom Heath and Christian Bizer, *Linked Data: Evolving the Web into a Global Data Space*, Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool, 2011.
[2] RDF 1.1 Concepts and Abstract Syntax, W3C Recommendation 25 February 2014, https://www.w3.org/TR/rdf11-concepts/.
[3] SPARQL 1.1 Query Language, W3C Recommendation 21 March 2013, https://www.w3.org/TR/sparql11-query/.
[4] Extensible Markup Language (XML) 1.0 (Fifth Edition), W3C Recommendation 26 November 2008, https://www.w3.org/TR/2008/REC-xml-20081126/.
[5] Batman Begins, 2005, https://viaf.org/viaf/176760887.

believe this generation, and the coming generation of digital humanities scientists, will need this in their DNA because they will be facing challenges that might require you to get a slightly more thorough understanding of, and control over, the systems they rely upon. Digital Humanities are no longer a mere compenetration of two communities serving the needs of one another. As a matter of fact, the humanities side of it is stimulating research questions, even for us computer scientists. That's why I want to insist on this.

So to quickly talk about myself like Rafael [Brundo Uriarte] did,[6] I also am a computer scientist at heart. Since my master's thesis work, I have been fascinated by the Semantic Web and, specifically, the topologies of ontology networks and how to optimise them for cognitive computing aspects like the ones that pertain to logical reasoning. Most of my postdoctoral work, however, has been about the application of these well, the principles and technologies and the research to several domains. I have done projects on Learning Analytics, on Smart Cities and on Industry 4.0, but what really captured my interest were the ones in the humanities, for this exact reason that I just mentioned and that I happen to have relayed just two days ago in Cambridge, that it's no longer one community serving the needs of the other, but as a matter of fact, both stimulating respective research questions from their own problems. Andrew has already mentioned LED, the Listening Experience Database,[7] which was actually my foray into digital humanities, from when I had just been hired by the Knowledge Media Institute at the Open University.[8] My line manager then said to me, 'Oh, by the way, we have this AHRC-funded project with the Royal College of Music[9] and the guy who was supposed to work on it left. Would you like to take it?'. Who was I to say no? The goal there was indeed to work together with the Royal College of Music, to build a catalogue of experiences of hearing music as documented in literature, being actual literature or official papers, correspondence, diaries, anything published but unsolicited. This bore several challenges for us as the computer scientists of the project team: (i) We had to capture a model human experience arising from aesthetics, intangible cultural heritage that wasn't being done very much at all. (ii) We wanted to do it natively as linked data, without transforming a relational database into linked data later. We were indeed the people with the hammer there, but we were not just seeing the problems as mere nails. Not only did we know how to work with Linked Data, but we also believed in its mission and wanted to make it that way. Not only that, but the database was crowdsourced: we had several dozens of users who were to actually input the data, and there was an editorial and curatorial workflow taking care of the quality of

[6] https://orcid.org/0000-0003-0750-7376.

[7] Alessandro Adamou, Simon Brown, Helen Barlow, Carlo Allocca and Mathieu d'Aquin, 'Crowdsourcing Linked Data on listening experiences through reuse and enhancement of library data' in *International Journal on Digital Libraries*, 20: 1, 2019 (61-79).

[8] Knowledge Media Institute, http://kmi.open.ac.uk.

[9] Royal College of Music, http://rcm.ac.uk.

those data. And somehow we had to shoehorn this into a system based on triple stores,[10] and RDF wasn't made for this. Very importantly also: yes, there were the technical principles and underpinnings, but the political aspect of it still wasn't covered (FAIR principles[11]). Back in 2013, implementing a system like that when we had no dedicated data publishing platforms meant having to do a lot of custom coding, taking an off-the-shelf content management system as was Drupal[12] – we didn't know by the way that WissKi[13] existed, but it would have helped us only so much. There is a lot of customisation required to make a Drupal installation talk to an RDF store, more so if you have to adapt the data management logic to support multi-tenancy over data, with many users having their say on a certain statement and then an overarching user approval of them.

Because I was a founding member of an Apache project called Stanbol,[14] which lasted from 2010 to 2020, and it was about actually bringing semantic services to content management systems, we used that technology to implement natural language processing and also lookup of third party linked data. Back then we didn't have Wikidata,[15] so we had to fall back to a less precise but still very powerful DBpedia that we were using alongside the British National Bibliography data, [16]

[10] By triple store, we intend the class of data storage solutions that adopt, as an atomic unit of data, a triplet of terms in a subject-predicate-object paradigm. Systems that allow the storage of data as quadruples, as is the case of multi-graph stores, are commonly referred to as quad stores.

[11] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao and Barend Mons, 'The FAIR Guiding Principles for scientific data management and stewardship' in *Scientific Data*, 2016.

[12] Drupal, https://www.drupal.org/.

[13] Martin Scholz and Guenther Goerz, 'WissKI: A virtual research environment for cultural heritage' in Luc de Raedt, Christian Bessiere, Didier Dubois, Patrick Doherty, Paolo Frasconi, Fredrik Heintz and Peter J. F. Lucas, *ECAI 2012 – 20th European Conference on Artificial Intelligence*, Montpellier: IOS Press 2012 (1017-1018).

[14] Apache Stanbol, http://stanbol.apache.org (retired).

[15] Denny Vrandecic and Markus Krötzsch, 'Wikidata: a free collaborative knowledgebase', in *Communications of the ACM*, 57: 10, 2014 (78-85).

[16] British National Bibliography, http://bnb.bl.ac.uk/.

MusicBrainz (through the Linked Brainz project)[17] and the VIAF[18] authority files. All this was then being synched with another linked data store, which was the one of my institute that was publishing all the linked data of all the projects and the organigram and the structure and everything of the Open University:[19] lots of work to address a grand challenge. And like I said, we were doing this based only on the technical principles, with only basic technological support.

Given the previous contributions, I realise you already have pretty good knowledge of what's behind the Linked Data principles, and also of the five-star open data model[20] that accompanies those principles. Both are brain-children of Sir Timothy Berners-Lee, father of the Web. What we lacked at the time of LED was the relationship between linked data and open data and the principles of FAIR data publishing: making your data findable, accessible, interoperable and reusable. The FAIR data principles were the first ones to actually have received backing and support from policymakers. This basically means you can actually put that on your funding bids to increase, if only slightly, the chances of them being accepted. That aside, it is still very important for us. Normally you would expect the political principles to come first: instead, this time we have had firstly technical principles coming up since 2006 all the way through for a good decade before 2016, when the paper describing FAIR data came out and happened to include several researchers from the Semantic Web community, and not by chance. The story today is significantly different, for even only the fact that the FAIR principles exist, and therefore many institutions have been striving to make it happen. Why? Because it so happens that Linked Open Data is a pretty good approximation of how you would implement the FAIR principles when publishing your research data.

Now, anyone who has worked on this knows, but usually won't tell you, that the relationship between linked open data and FAIR is the apex of 'easier said than done'. There is just no one-size-fits-all solution that allows you to make FAIR data out of anything, so you need different approaches. As a matter of fact, when I joined the Hertziana, I came across the realisation that there was so much heterogeneity, along with so much inherent value in decades worth of research projects and catalogue endeavours having been carried out on different levels in different departments and with varying funding. There are several levels, like the level of the actual digital library and photographic library and the cataloguing projects that go alongside it. You've heard about the Fotothek, the library, of course, and its association with the Kubikat project. Other important projects include the

[17] LinkedBrainz, https://wiki.musicbrainz.org/LinkedBrainz.
[18] Virtual International Authority File, http://viaf.org/.
[19] Enrico Daga, Mathieu d'Aquin, Alessandro Adamou and Stuart Brown, 'The Open University Linked Data - data.open.ac.uk' in *Semantic Web*, 7: 2, 2016 (183-191).
[20] 5 ★ Open Data, https://5stardata.info/en/.

ZUCCARO database of art history in Rome[21], which is founded on principles like the Semantic Web ones, but with a different implementation; the CIPRO catalogue of maps[22], and then several research-specific projects, as it were, of which I'm going to give a few examples, because I've been trying to get those to join the holistic knowledge graph of the Hertziana.

Making digital humanities data FAIR through Linked Open Data: scenarios and approaches

It is not one single technical challenge being discussed here. It is, as a matter of fact, a family of problems arising from the differences in how the data-intensive research projects at the Hertziana were conceived. There may have been projects that simply ended up dead in the water when funding ran out, or when the person responsible left and nobody would maintain it. However, there is still, very importantly, an interest in the content of those projects because, unlike many projects of computer science and hard sciences, data in the humanities and arguably in art history hardly become obsolete. If anything, once they become too old, we talk about them in terms of historiography, but hardly any of the data at the centre, or the process behind them, can be considered obsolete. The technology may, however, become outdated, whilst the lifetime of research projects might be affected by them being born to respond either to specific research questions, or to more general questions. As an occurrence of the latter, 'let's build a catalogue of this, let's build a catalogue of that' is commonplace in the humanities, and LED was exactly that: 'let's build a catalogue, then we'll get another grant for doing specific studies on it', which was exactly what happened.

Not every topic is vertical, but usually, when you have a vertical topic to address, you want to implement the technological stack in the quickest way you can. So, typically, project leads will just be like: 'okay, give me a content management system like WordPress[23], a relational database, we get the thing done and it's out the door quickly' – by the way, it isn't. And then, of course, there is a need to build the culture of open data, an understanding of what it means to actually licence your data as you would your software, and that licensing your data is only part of the problem. Licence goes side by side with the attribution of data, which is even more critical than one would actually experience in e.g. open education or learning analytics. In a way, attribution and licensing replace, or rather develop, the old problem of copyright: still, it is a culture that needs to be built. All this diversity should be taken advantage of and turned into opportunities for, for example, giving a second life to projects that are still important yet only known by a few.

[21] Martin Raspe and Georg Schelbert, 'ZUCCARO—Ein Informationssystem für die historischen Wissenschaften', in *IT - Information Technology*, 51, 2009 (207-215).
[22] Catalogo illustrato delle piante di Roma online, http://db.biblhertz.it/cipro/.
[23] WordPress, http://wordpress.org/.

Figure 1: Cross-referencing of the Church of San Crisogono in Rome (above, left and right) within projects at the Bibliotheca Hertziana and (below) with external authorities.

As I was saying, not much in our data becomes obsolete: take, for example, the linking problem. In several databases hosted within the remit of the project at the Hertziana, we already have cross-links describing different entities. Figure 1 shows an example for the Church of San Crisogono in Rome, which I will carry forward as a running example through this presentation. On the left side, it is shown how San Crisogono appears in a project called LVPA, which will be discussed later. The LVPA XML code has a link to both the ZUCCARO catalogue that was mentioned earlier, and to the equivalent of San Crisogono in the GND[24] authority file. Similarly, there is another project called Roma Communis Patria,[25] which uses a separate database in the form of an online, seemingly old-fashioned yet effective tabular database, which links also its own data to ZUCCARO, but also to the maps of Nolli, which are catalogued by the CIPRO project. On the right side, we have another example of linking an aspect of San Crisogono, which is the relation to architect Giovanni Soria. Two external data sources are used: on the one end the GND again, and on the other the Getty list of artist names,[26] both linked in

[24] Gemeinsame Normdatei, https://d-nb.info/standards/elementset/gnd.
[25] Roma Communis Patria, https://www.biblhertz.it/it/roma-communis-patria.
[26] Union List of Artist Names (ULAN), https://www.getty.edu/research/tools/vocabularies/ulan/.

the Fotothek data using a specific schema: the MIDAS schema[27] which originated in Marburg. Both are completely non-standard ways to create references, rather, for things to reference one another. Neither XML code is not canonical, nor is the table: everything is bespoke, as modelled by the person responsible for the project at hand: I endeavour to do something about it, starting with, if you will, the easy example.

The first example is a legacy project which, in a way, we want to revitalise. The upside of having a so-called legacy project no longer being developed is that, if you can get your hands on it, you can pretty much do what you want with it, so long as you are not disrupting the workflow of the scholars who are still using it today. Luckily, many such projects leave a semi-structured trace behind them: a relational database, some emails, some HTML, even plaintext, which sometimes can be considered as structured. What one can do in this case is to actually build a transformation layer that can either be used to perform an ETL[28] process onto a triple store; or, for example, leave it persistently as a layer to perform virtual data integration in real time, if computationally possible, with the other available sources. There are several tools and software libraries for either option; even some that are simply fed some data and they just go their own way, silently creating the Linked Data version – the RDF code – out of it. Examples include the Apache any23,[29] R2RML,[30] which is specific to relational databases, although it's a standard recommendation; the X3ML engine,[31] which is heavily used in the first project already mentioned, and is used to systematically transform large chunks of XML data with huge mapping definitions into RDF.

There is, however, another approach, which mandates that you don't necessarily have to transform everything in one take: you should instead be able to query non-semantic, non-RDF data, as you would query any data on the fly. Lately, this has been the leading research approach, where a few actors in the field are trying to somehow push towards virtual data integration. SPARQL-Generate[32] is one example, actually: it's a dialect of the SPARQL language, which is standard for querying data used specifically for defining primitives, to create RDF data. And

[27] Jens Bove, Lutz Heusinger and Angela Kailus, 'Marburger Informations-, Dokumentations- und Administrations-System (MIDAS): Handbuch und CD' (Literatur und Archiv; 4).
[28] Extract-transform-load
[29] Apache any23, https://any23.apache.org/.
[30] R2RML: RDB to RDF Mapping Language, W3C Recommendation 27 September 2012, https://www.w3.org/TR/r2rml/.
[31] X3ML Engine, https://github.com/isl/x3ml.
[32] Maxime Lefrançois, Antoine Zimmermann and Noorani Bakerally, 'Flexible RDF generation from RDF and Heterogeneous Data Sources with SPARQL-Generate', in Paolo Ciancarini, Francesco Poggi, Matthew Horridge, Jun Zhao, Tudor Groza, Mari Carmen Suarez-Figueroa, Mathieu d'Aquin and Valentina Presutti, *Knowledge Engineering and Knowledge Management - EKAW 2016 Satellite Events*, Bologna: Springer 2016 (131-135).

then Facade-X[33] is one being developed by my former colleagues at The Open University: it is actually quite interesting because it strives to be literally about querying anything, pretty much by using nothing but the SPARQL language; not a dialect of SPARQL, but the language itself, which is then compounded with additional utility functions for manipulating strings, creating entities, and most importantly, accessing your local data as if accessing external SPARQL services on the Web.

```
<poi>
   <title>San Crisogono</title>
   <variant/>
   <coordinates>12.473248,41.889052</coordinates>
   <notes>TCI 561-562</notes>
   <equivalent domain="zuccaro">
      <title>S. Crisogono</title>
      <id>1235</id>
   </equivalent>
   <equivalent domain="gnd">
      <title>San Crisogono in Trastevere</title>
      <id>4461207-2</id>
   </equivalent>
   <recurrence image="yes">
      <year>1901</year>
      <quote>Sancti Chrysogoni ›San Crisogono‹ ◉</quote>
      <cite>Forma Urbis Romae  (1901) [18]</cite>
      <link>https://universalviewer.io/uv.html?manifest=https
   </recurrence>
   <recurrence image="no">
      <year>1891</year>
      <quote>›San Crisogono‹</quote>
      <cite>Le chiese di Roma dal secolo IV al XIX (1891) [70
      <link>https://universalviewer.io/uv.html?manifest=https
   </recurrence>
 </poi>
```

Figure 2: XML snippet describing the Church of San Crisogono in Rome.

```
CONSTRUCT {
  ?x rdfs:label ?title
    ; a crm:E22_Human-Made_Object , gndo:BuildingOrMemorial
    ; skos:altLabel ?altLabel
    ; crm:P53_has_former_or_current_location ?loc
    ; owl:sameAs ?zuccaro , ?gnd
  .
  ?loc a crm:E53_Place
    ; crm:P168_place_is_defined_by ?wkt
} WHERE {
  SERVICE <x-sparql-anything:file:///Users/adamou/workspaces/biblhertz/ld/lvpa/data/lupa-export.xml> {
    [] a xyz:data ; fx:anySlot ?poi .
    ?poi a xyz:poi ; fx:anySlot [
            a xyz:title ; rdf:_1 ?title ].

    OPTIONAL { ?poi fx:anySlot [
            a xyz:variant ; rdf:_1 ?altLabel ]}

    OPTIONAL { ?poi fx:anySlot [
            a xyz:coordinates ; rdf:_1 ?coordinates
    ] FILTER(?coordinates != ",") }

    OPTIONAL { ?poi fx:anySlot [
            a xyz:equivalent ; xyz:domain "zuccaro" ;
            fx:anySlot [ a xyz:id ;
              rdf:_1 ?id_z ] ] FILTER (strlen(?id_z)>0) }
  }
  BIND (fx:entity( data:, "builtwork/lvpa/", ?uuid ) AS ?x)
  BIND (fx:entity( ?x, "/loc/", fx:serial(?x) ) AS ?loc)
  BIND( STRDT( CONCAT( "POINT(", REPLACE(?coordinates, ',', ' '), ")" ), geo:wktLiteral ) AS ?wkt )
  BIND (fx:entity(data:, "builtwork/zuccaro/", ?id_z) AS ?zuccaro)
}
```

Figure 3: Generative SPARQL query for Facade-X to convert the data from Figure 2 to RDF.

[33] Enrico Daga, Luigi Asprino, Paul Mulholland and Aldo Gangemi: 'Facade-X: an opinionated approach to SPARQL anything', in *CoRR* abs/2106.02361, 2021.

Figures 2 and 3 show an example of how we can query some resources on churches represented in the LVPA project. LVPA is, for all intents and purposes, a gazetteer which connects the toponyms found in it with references to whichever works present in the Bibliotheca Hertziana collections of rare books that mention them or describe them in any way. Figure 2 shows a slightly expanded example of how San Crisogono is organised and linked to in LVPA. There are a couple of ways to query all this directly in SPARQL like it were an RDF database: Figure 3 shows how this can be achieved using the Facade-X library called SPARQL-Anything.

```
<http://data.biblhertz.it/builtwork/lvpa/5fdaf203b0f47ab70c6aa19e87558b06> a crm:E22_Human-Made_Object,
        gndo:BuildingOrMemorial ;
    rdfs:label "San Crisogono" ;
    crm:P53_has_former_or_current_location <http://data.biblhertz.it/builtwork/lvpa/5fdaf203b0f47ab70c6aa19e87558b06/loc/1> ;
    owl:sameAs <http://data.biblhertz.it/builtwork/zuccaro/1235>,
        <https://d-nb.info/gnd/4461207-2> .

<http://data.biblhertz.it/builtwork/lvpa/5fdaf203b0f47ab70c6aa19e87558b06/loc/1> a crm:E53_Place ;
    crm:P168_place_is_defined_by "POINT(12.473248 41.889052)"^^<http://www.opengis.net/ont/geosparql#wktLiteral> .
```

Figure 4:RDF output resulting from the application of the query on Figure 3 to the data of Figure 2.

First of all, the one in Figure 3 is very close to standard SPARQL code that doesn't just perform a selection query, a projection of the data. It's a CONSTRUCT SPARQL query: in SPARQL, as some of you know, the CONSTRUCT clause is used to generate new data out of query results. This is the procedure: spit out new RDF triples out of the solution of a query to the dataset, then apply the Facade-X approach. This gives an abstraction over the XML data. So therefore one sees the tags, the attributes, even the text within those tags as if they were RDF lists, which can be queried normally as any RDF list can, usually via the RDF underscored notation (e.g. rdf:_1, rdf:_2 and so on) or special predicates made available by the Facade-X library, like this fx:anySlot. Then, there are functions one can use to generate URIs, literals and, more in general, RDF nodes from all this. Another important feature of Facade-X is that it allows you to also work with the service definition: normally, this is used to query the endpoint of an external dataset, but I am using it here to query a local file: if it were being served on the Web, we could query the online version directly. The result looks like on Figure 4, which is in the serialisation form called RDF/Turtle and basically realises the application of the query. Basically it has transformed a chunk of what was in that XML tree into a series of RDF statements that tell us that this point of interest corresponds to a Human-Made_Object and it happens to be a building; that it is called San Crisogono; that it matches an entity in the ZUCCARO database and one in GND; and that its location, which has its own URI, has certain coordinates in a standard geospatial literal format, so that it can actually be pinpointed. This is a totally standard representation form that anyone who has a working knowledge of

CIDOC-CRM[34], some FRBR[35] and the basic RDFS meta-model of ontologies can work out easily.

The Fotothek[36] is a slightly different story because that project is alive and running, in which case we need to proceed slightly differently. We have an interface service that somehow compounds the human-facing side of things and we can use IIIF[37] services and Web APIs[38] to query, if not for the San Crisogono entity, at least for the string San Crisogono: that's already something. And then we combine that capability with the work that is being carried out in the PHAROS consortium[39] to use the X3ML library to convert the whole bulk present in the XML dump of the Fotothek data into RDF. We can therefore obtain standard CIDOC-CRM-based RDF as spat out by X3ML, which tells us, for example, that San Crisogono is composed of a few more architectural structures, such as the *campanile* – the bell tower, the crypt, and of all the actual physical objects for which there is a corresponding digitised photo, for instance, the pictorial cycle of the lives of the Saints Benedict and Sylvester. By these means we can also obtain their RDFS labels, if only in German. The one of labels in the photos being almost entirely in German is, by the way, a separate issue and another job we need to actually undertake.

There is then another scenario i.e. the case where the project is still running, there are people working on it but, for reasons of quickness, of expertise, or of time-to-publish, employ a technological stack that is incompatible with the Linked Data ones. This is to no detriment to the project managers: it is in fact a completely legitimate choice, the point then being, that we should find ways to let them keep working on their tabular system, WordPress static pages and what have you, if so they wish, whilst providing an alternative editorial and curatorial workflow to manage the same data in a semantic way, with the expectation that they will eventually embrace it. That is the case of a project within the Hertziana called Roma Communis Patria, built upon a cloud-based tabular database and a content management system. There are tools for dealing with such cases too. Typically, one will still perform an integration process to create a linked data counterpart, i.e. an RDF graph, of what they're working on with their CMS + tabular database combination, in order to provide an alternative version of it that can update itself and seamlessly integrate itself with the others. The goal is therefore to show project leads the benefits of the process: 'Look what we can do: if you convert your data in

[34] Martin Doerr, 'The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata', in *AI Mag*, 24: 3, September 2003 (75-92).
[35] Olivia M.A. Madison, 'The IFLA Functional Requirements for Bibliographic Records', in *Library Resources & Technical Services*, 44: 3, June 2000 (153-159).
[36] Bibliotheca Hertziana Photographic collection, http://foto.biblhertz.it/.
[37] Internet Image Interoperability Framework, https://iiif.io.
[38] The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), https://www.openarchives.org/pmh/.
[39] PHAROS: The International Consortium of Photo Archives, http://pharosartresearch.org/.

this way and need to link to other institutional databases (in our case, ZUCCARO and LVPA being examples), that's the way you can reach out to them'; then offer continued support to them, possibly with a virtual research environment.

```
        SELECT DISTINCT ?church ?churchLabel ?place ?placeLabel ?location ?dist WHERE {
 VALUES(?church ?churchLabel ?coords) {
  ( bldg:NB02 'Pontificio Collegio Nepomuceno'
              'POINT(12.509671697349775 41.878010732109146)'^^geo:wktLiteral )
            ( bldg:NB04> 'San Basilio agli Orti Sallustiani'
              'POINT(12.490608009652812 41.90506747768336)'^^geo:wktLiteral )
 }
 ?place wdt:P31/wdt:P279* wd:Q24398318         # Is a religious building
  ; wdt:P31 ?instance
 SERVICE wikibase:around {
    ?place wdt:P625 ?location .
    bd:serviceParam wikibase:center ?coords .
    bd:serviceParam wikibase:radius '.1' .          # within 100 metres
 }
 SERVICE wikibase:label { bd:serviceParam wikibase:language 'it'. }
 BIND(geof:distance(?coords, ?location) as ?dist)      # compute distance
} ORDER BY ?church ?dist
```

Figure 5: Federated SPARQL query to Wikidata for the retrieval of religious buildings within proximity.

Achieving the above takes quite some procedural code: one cannot expect to be able to perform bespoke integration entirely, with only a transformation layer based on query languages. Oftentimes, query languages do not allow looping structures to be created easily, therefore iterations can be a problem. Usually, when having to iterate over something, you need to adopt procedural code. It is not as tough as one might think it is: Figure 5 shows an example SPARQL query which is executed by procedural code in Python language that an actual Digital Humanities intern at the Hertziana worked on over the course of a few months. The goal of this query is to create links between churches in the Roma Communis Patria project and Wikidata. This was achieved with the parametric SPARQL query in Figure 5, which is sent to the Wikidata endpoint. The VALUES clause is an embedded table that samples two building of the database, out of about sixty present. Given the names and the coordinates, the query says: 'Find anything that is a religious building – which includes churches, religious colleges and abbeys – that lies within 100 metres from here, and retrieve their labels'. Then, in the procedural code, we can perform string matching to find the most likely matches, using off-the-shelf matching algorithms, and then picking the best match and linking the corresponding URIs. Here, querying Wikidata does a large portion of the job for us.

Virtual research environments

Only a few years ago, bootstrapping a new project based on LOD would have been inconceivable. However, now we have virtual research environments (VREs), which are often themselves based on Linked Data. VREs are, in a strict sense, those software systems that support the life-cycle of research data, possibly in a collaborative way: usually the criteria, workflow and sometimes publishing are supported; these systems can be content management system in their own right, like the traditional WordPress and Drupal and extensions thereof. There are entire Wiki-based systems with simple methods for creating templates out of the data; data patterns, in fact, that can be extracted by simply looking at the properties that link the data. Some of these systems are natively based on Linked Open Data: ResearchSpace,[40] developed by the British Museum, is entirely based on triple stores. It supports the CIDOC-CRM model in its workflow and allows you to create templates not only for displaying the data, but also for inputting them. If I had had this at the time of LED, it probably would have saved me several months of work trying to encode Linked Data, inputting for users, making sure they wouldn't make too many mistakes, like not linking data where they should, or linking to the wrong ones. Other implementations include the one provided by Metaphacts,[41] mostly commercial, and then an implementation on top of its open source version, which is indeed spearheaded by the RDS project[42] of the Swiss Art Research Infrastructure consortium. The Bilder der Schweiz online[43] is probably the best known example in these cases. These systems have their own standardised SPARQL API and expose their own query endpoint: queries on a local SPARQL service can therefore be federated to those endpoints, so that their results are integrated with the solution to be computed for the local query. An example of another type of system is WissKi, based on the Drupal CMS, which is Wiki-like but backed by a triple store.[44] An integration strategy here could be to bypass WissKi, yet taking advantage of the fact that it does have a SPARQL point that can be queried directly, so that its results are integrated as shown earlier. Other systems don't really have a SPARQL API and are not based on a triple store, but they are still semantic systems. Omeka-S[45] is one of them: it has its own REST API for reading and searching and writing data. This is possible but definitely takes more work: having to convert SPARQL queries into lookups to the search API or calls to the read API is of course limited and can only

[40] Dominic Oldman, Diana Tanase, 'Reshaping the knowledge graph by connecting researchers, data and practices in ResearchSpace', in Denny Vrandecic, Kalina Bontcheva, Mari Carmen Suarez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, Elena Simperl, *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference*, Monterey: Springer 2018 (325-340).
[41] Metaphactory, https://metaphacts.com/product.
[42] RDS, https://rds.swissartresearch.net.
[43] Bilder der Schweiz online, https://www.bilder-der-schweiz.online/.
[44] Scholz and Goerz, 'WissKi'.
[45] Omeka-S, https://omeka.org/s/.

be done for a subset of the SPARQL language. It all depends upon being able to get a hold of the structured traces that those systems leave: if they don't even have these, they are not proper virtual environments and probably shouldn't be taken into consideration.

By combining all these strategies, we can avail ourselves of a combined query engine which, through one service endpoint, represents the entire knowledge graph of an institution. This can, in theory, be achieved without any need for a dedicated triple store to serve data independently of third-party platforms, however, in the presence of ETL processes, having one becomes an advantage. It is important to note that integrated Linked Data systems have the flexibility to accommodate the presence or absence of dedicated triple stores.

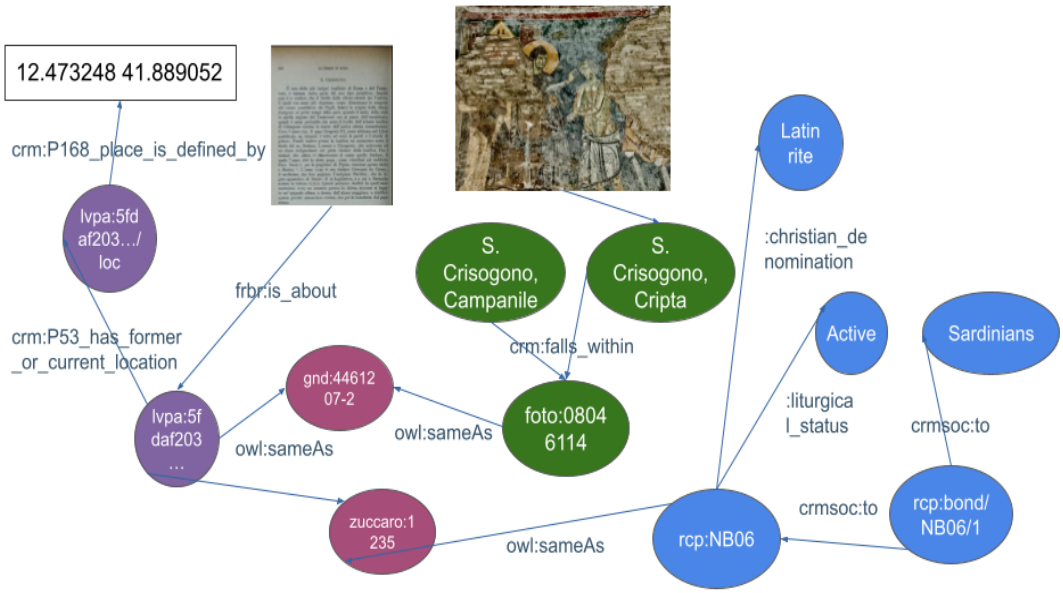

Figure 6: Integrated view of the Church of San Crisogono in the Bibliotheca Hertziana knowledge graph.

Back to the case of San Crisogono, we can then present the available knowledge about it in Figure 6, by combining the query output from different projects: LVPA, ZUCCARO, the Fotothek, and Roma Communis Patria all tell us different facts about this church. For one thing, LVPA tells us (purple nodes) that it has certain geographical coordinates and that there are publications mentioning the church. Then, the Fotothek tells us (green nodes) that there are monuments within the church complex, and that a number of works of art can be found within parts of the complex. Roma Communis Patria tells us (blue nodes) that, at some point in its existence, the church was assigned to the Sardinian *natio*, or community; that, despite still being an active church, it is no longer assigned to the Sardinians; and that it is of Latin rite. All the nodes that are linked by edges of type owl:sameAs, representing the equality property, can be conceptually collapsed to obtain one

holistic view of San Crisogono, as it is described by these four projects. But how does one make this linking happen? One possible way is a very simple boilerplate SPARQL query that uses (owl:sameAs|^owl:sameAs)+ as predicate and elects one of the possible URIs that identify the church – such as the GND for German institutions like ours – as object. Once sent to our unified endpoint, this query returns RDF data that state exactly what was said earlier: LVPA provides a link to GND and to publications mentioning the church, for which there is a link to the corresponding IIIF manifest; further on, ZUCCARO provides the GeoSPARQL point coordinates of its location; Roma Communis Patria describes the Christian denominations and assignment to the Sardinians, that the aforementioned location happens to be in the *rione* Trastevere and has an address in Piazza Sonnino: everything brought together. It is even possible to collapse these URIs into one by slightly manipulating the SPARQL query, so that everything appears to have one subject, though the approaches are equivalent from a logical perspective, as all the URIs denote the same thing.

A new paradigm emerges: Linked Open Usable Data

We have illustrated techniques to make project data FAIR, particularly embracing the Linked Open Data implementation of FAIRness, in ways that vary depending on the scenario being faced. Such scenarios, however, need to be regarded from lenses that are not only technical, but also procedural, curatorial, or administrative. There are ways to take many of these factors into account when making data FAIR. It is, at this point in time, appropriate to mention why the title mentioned LOUD instead of LOD. LOUD is actually a more recent paradigm which happens to have been defined in cultural heritage and in the Humanities. LOUD, as in Linked Open Usable Data,[46] is an acronym that was presented first in 2018 by Rob Sandersonof the Getty Trust.[47] Similarly to the five-star open data model, they are trying to build a grading scale of what usability means when one is talking about open data. The 'star system' of LOUDness may not appear as precise as the one of open data, but it offers an interesting insight, because none of the usability considerations of the LOUD grading scale has to do with human-computer interaction, or with user experience (UX). It is primarily concerned with the morphology of the data, the ontologies they are based upon and how these should be documented. This includes: (1) 'The right assumption for the audience', i.e. basing the generation of data upon use cases that drive them, because these eventually will somehow guide how data will be modelled according to certain schemas, how those schemas are going to be documented, and how the data can be synthesised for the benefit of those using them for specific research questions. (2) 'Few barriers to entry: If it takes time to understand the model, the query, syntax and so forth', people will move to other technologies or services ('look for easier targets'). (3) 'Comprehensible by introspection': you shouldn't be having to study the ontologies, but you should be

[46] Linked Open Usable Data, https://linked.art/loud/.
[47] https://orcid.org/0000-0003-4441-6852

able to understand the model you're looking at without having to study them. Finally, (4) 'Documentation with working examples' and (5) 'Few exceptions, many consistent patterns' recommend that data stewardship should work more towards the pattern than towards the exceptions to those patterns.

Whether and how the introduction of VREs helped the realisation of the LOUD paradigm is a legitimate question. For one, VREs do a great job towards the fifth star, i.e. consistent patterns, because they force the user/developer to think in terms of templates. This makes it relatively hard to deal with exceptions unless one creates bespoke templates for handling them, and that would in turn defeat the very purpose of using VREs in the first place. The second star, i.e. minimising the barriers to accessing those data, is somewhat guaranteed by the fact that VRE managers will be creating their own way of documenting the data with the ontology underneath, as an alternative to asking users to explore the ontology if they are not confident doing so. As for providing the right abstraction, comprehensibility by introspection, and working examples: whilst undoubtedly facilitated by VREs to an extent, these still require a great deal of work by the maintainer of the VRE instance to implement these properly. Awareness of the use case at hand helps even when one has little expertise of standard ontologies such as CIDOC-CRM, FRBRoo and the like.

To summarise, a LOUD-based architecture is definitely possible today: it requires combining several components and several transformation workflows at the same time, depending on how many heterogeneous approaches are being dealt with. Still, this is becoming possible even at the interoperability level of services. Rafael [Brundo Uriarte] gave a very good example of it when he was discussing microservices and how one would orchestrate services from different institutions: that is, as a matter of fact, a very good way to consume the data other than having to query them by SPARQL or having to understand what comes out of looking up the URI on a Web client (browser or other). The added lesson learnt here is that all this can be made to work together with traditional relational databases, without having to take them out of the picture completely, by adding a layer of transformation primitives. The fact that the Getty Trust felt the need to come up with this LOUD paradigm somehow reflects that: implementing, understanding and using open data is still easier said than done. LOUD covers aspects of the FAIR principles that have ties with reusability and accessibility and, to a lesser degree, findability and interoperability. It is by no coincidence that these principles originated in the Humanities, as did most VREs illustrated today like ResearchSpace, WissKi and Omeka-S: many of them were indeed born in the context of humanities projects. The heterogeneity of research projects tends to complicate a uniform usage of VREs, but at least it is now possible to fulfil use cases on an operational level, which only ten years earlier could not be carried out efficiently.

Conclusion

On a concluding note, the goal of using rich research environments should not cause someone to become dependent on this or that environment: we must be wary of

mitigating another risk of vendor lock-in because a certain VRE was adopted and its developers went bust, leaving us with no way of publishing or using those data. This is no longer the case, as they are still semantic data in RDF using ontologies that are very common. That alone means you can seamlessly migrate to another environment tomorrow, and even if you don't, the data will still be readable, understandable by humans and machines alike, even in a decade or, hopefully, in five hundred years.

Alessandro Adamou is a Digital Humanities Scientist at the DHLab. His main expertise is on Data Science, primarily in the areas of the Semantic Web, Knowledge graphs, ontology networks and their cognitive computing applications in the Humanities. One of his core research interests is on finding ways to formally model intangible aspects of cultural heritage, such as experience or tradition. With the Royal College of Music, Alessandro has developed LED, the first crowdsourced linked dataset on evidence of listening experiences; he has also collaborated with the Irish Traditional Musical Archive for the LITMUS project and contributed knowledge modelling of reading experience for the READ-IT project. He is active in several outreach endeavours of data science for the Humanities, such as being regular organiser and co-chair of the WHiSe workshop series (2016-present). He has supervised doctoral research work on studying cultural contact in early Roman Spain through Linked Data. Previously, Alessandro has held researcher position at Insight SFI Research Centre for Data Analytics (Galway, Ireland), The Open University (Milton Keynes, UK) and the National Research Council (Rome, Italy), where he explored transversal applications of Linked Data to several other domains, such as Learning and Education, Smart Cities (for which he has won several awards and nominations), Content Management, and Industry 4.0.

alessandro.adamou@biblhertz.it