10 Linked data strategies for conserving digital research outputs

The shelf life of digital humanities

Florian Kräutli, Esther Chen, and Matteo Valleriani

Introduction

The digital humanities struggle with the challenge of long-term usability and accessibility of digital research outputs. Research data and digital artefacts in the form of databases and websites constitute essential results of digital research projects, yet they are typically not maintained in ways that can be reused, cited and accessed in the long term. Our main questions are, therefore: how can we maintain digital humanities research outputs so that they remain accessible and usable? What requirements must the research data infrastructures of cultural heritage institutions meet in order to fulfil this task? How can research outputs be linked to the primary sources being studied and how closely do central library infrastructures and individual research projects have to be aligned?

These questions have been posed for some time now. Since the early 2000s, the European Union has established large funding programs to develop transnational research infrastructures in various disciplines. This funding aimed to address these open questions and to increase the development and competitiveness of Europe as a research space. The projects included in the term "research infrastructures" diverge as widely as the expectations and hopes for them, especially in the humanities. In her paper "What are research infrastructures?", Anderson (2013) examines a wide range of research infrastructures, taking into account big funding programs such as the European Strategy Forum on Research Infrastructures as well as many different national projects such as the French Biblissima project or Oxford's Cultures of Knowledge. She emphasises the role of research infrastructures as an experiential presence that is embedded in the practices and experience of research, claiming that the strong collaboration of scholars, librarians and archivists is a major key to success (Anderson 2013, 10):

> Infrastructure development and take up is far more successful if it emerges from researchers' own practices: if it fills gaps in existing provision, or it is a solution to identified problems and perceived difficulties.

In her book Scholarship in the Digital Age: Information, Infrastructure, and the Internet (Borgman 2007), Borgman underlines the significance of

DOI: 10.4324/9781003131816-19

research data while examining what she calls information infrastructures. She claims that no framework exists for research data comparable to that for publishing, while at the same time the output of such data increases rapidly. Looking back at this statement from today's perspective, we still cannot regard this problem as being solved, even though digital research data is deeply embedded in the day-to-day practices of humanities research.

At the Max Planck Institute for the History of Science, we draw from a comparatively long history of digital scholarship, notably through the ECHO initiative that began to digitise historical sources and publish them on the Web as early as 2002 (Renn 2002). Since then and to date, research projects have been studying, annotating and contextualising digital sources, which have become increasingly available in recent years. In addition to common scholarly outputs such as books or journal articles, these projects produce digital outputs such as websites, databases or virtual exhibitions: typical artefacts of digital humanities research. Maintaining these artefacts has, however, proven to be challenging. Unlike their physical counterparts, e.g., books, monographs and journal publications, they do not end up in the library and instead live on scattered servers and ageing software systems. This makes maintaining long-term access to these resources difficult and ensuring that they are usable and interoperable with evolving digital technologies is nearly impossible.

Our ambition is to complete the digital research life cycle: to make sure that digital research outputs can be discovered, accessed and reused within one integrated environment. We seek to achieve this by adopting a common model to represent our digital knowledge and by implementing Linked Data technologies for data storage and exchange. In this chapter, we outline the challenges surrounding the preservation of digital humanities research outputs and present how we address them, both within the scope of an individual research project and at scale, as libraries take on new responsibilities in managing digital research outputs. First, we outline some of the main challenges in the preservation of digital humanities research as identified by the scholarly community. We then present two of our projects as case studies in which we tackle these challenges. We propose to look at digital humanities research outputs as consisting of two layers: a presentation layer and a data layer. We suggest that there is a need to focus on the data at the expense of its presentation if we are to seek to preserve these research outputs in reusable and sustainable ways.

Current challenges in maintaining digital humanities research outputs

The challenge of creating reusable data

With the increasing adoption of digital methods, the need for reproducibility of research is no longer confined to the natural sciences but has become relevant for the humanities too (Peels and Bouter 2018). As O'Sullivan (2019) writes: "Humanities scholars are increasingly expected to accept the findings of their peers without access to the data from which discoveries are drawn. Access to data is just part of the problem." The other part of the problem that O'Sullivan discusses is a lack of documentation and transparency about the applied methods, which often make reproducing digital humanities research impossible. We prefer to keep our focus, however, on the problem of access to data. Beyond mere access, the problem lies in making and keeping data usable, i.e., "not to archive data, but to keep them alive" (Kilchenmann, Laurens, and Rosenthaler 2019). Thorough data curation is therefore not only a necessity, but due to the richness of humanities data, also a substantial challenge (Henry 2014). Archives and libraries employ a range of standard models to describe their holdings in a common and reusable way, such as Machine Readable Cataloguing (MARC, MARBI, 1996), Encoded Archival Description (EAD, LOC, 2002) or Bibliographic Framework (BIBFRAME, LOC, 2016). While these models allow for rich descriptions of digital collections data, they are not necessarily compatible with each other, which complicates data sharing and reuse. In addition, although they are meant to be broadly applicable, they may not support research in the wider framework of humanities research. As Oldman et al. (2014) write: "Cultural heritage data provided by different organisations cannot be properly integrated using data models based wholly or partly on a fixed set of data fields and values, and even less so on 'core metadata'".

Scholars may not only need to describe a wide range of material but also events, actors and the relationships between them, as well as observations, conflicting information, beliefs and inferences. The conceptual reference model CIDOC-CRM[1] was developed to address the problem of incompatibility between standards (Doerr and Crofts 1999; Crofts et al. 2011) and to allow the description of humanities data to a high level of accuracy. It has therefore emerged as a general-purpose model for the cultural heritage domain (Oldman, Doerr, and Gradmann 2016). CIDOC-CRM defines a basic set of entities such as actors, places, concepts and, most importantly events. This approach allows for things to be described not through a vocabulary of terms, whose meaning might be ambiguous, but through events that create or transform things and can happen at a particular place or time and through actors. In comparison to existing approaches, these generic and minimal building blocks allow for a "less complex, more compact and sustainable model, but with far richer semantics" (Oldman, Doerr, and Gradmann 2016). Instead of using potentially ambiguous terminologies such as "author", e.g., the model allows for a detailed digital representation of events that lead to the creation of a particular cultural artefact along with relationships to the actors involved.

[1] CIDOC stands for the International Council for Documentation, under whose patronage the CIDOC-CRM Special Interest Group maintains and develops the reference model.

Extensions of CIDOC-CRM have been and continue to be developed where a consensus about the material and events that are represented allow for greater specificity of the data model. This includes FRBRoo (Doerr et al. 2013; Bekiari et al. 2015), a CIDOC-CRM extension of the bibliographic standard FRBR (the "oo" in FRBRoo stands for "object-oriented"; Functional Requirements for Bibliographic Records, IFLA 1998) and CRMdig (Doerr, Stead, and Theodoridou 2016) for capturing digitisation and provenance information. Each of these extensions introduces new entity types, which are derived from the basic set specified in CIDOC-CRM but cater to the individual needs of their subject domains. CIDOC-CRM provides a generic class for describing physical carriers of information, which FRBRoo extends to allow for distinguishing specific types of information carriers such as printed books, audio CDs or videotapes.

The absence of suitable interfaces and platforms for the user-friendly creation of semantically rich data according to the CIDOC-CRM model has previously been a major obstacle for their adoption. By now, a growing number of solutions have been created, e.g., WissKi (Goerz et al. 2009), ResearchSpace (Oldman 2016) and Metaphacts Open Platform (Metaphacts 2019). These platforms allow researchers to create semantically rich data according to the CIDOC-CRM model without necessarily having to be familiar with all its intricacies.

Besides the conceptual representation of data according to a standard such as CIDOC-CRM, the file format in which data is ultimately stored is a deciding factor in its reusability. As the PARTHENOS project, an EU-funded initiative for enabling interoperable digital humanities research, states: "There will never be one standard format for all data. Rather, we must find means to translate between them" (PARTHENOS, n.d.). A file format that facilitates translation across data schemas, as well as interlinking of data, is the Resource Description Framework (W3C 2014). In this format, "Meaning is expressed by RDF, which encodes it in sets of triples, each triple being rather like the subject, verb and object of an elementary sentence" (Berners-Lee, Hendler, and Lassila 2001, 38). Due to its flexibility, RDF has emerged as common ground on the data format level to create, preserve and exchange digital humanities research data. The Swiss Data and Service Center for Humanities (DaSCH) caters to a variety of needs in humanities research through an RDF-based data infrastructure (Kilchenmann, Laurens, and Rosenthaler 2019): "RDF allows great flexibility of data modelling, which enables the DaSCH to use one single infrastructure for data, metadata, models and structures for any project regardless of the data concept used. Thus, the DaSCH has to maintain only one single infrastructure to provide sustainability. Data from any one project can be analysed and compared with data from other projects".

RDF is also a central building block of the Semantic Web. The Semantic Web and the application of Linked Data principles seek to make data reusable by focussing on machine readability and dense interconnectedness through the technical principles of the Internet. As a concept, the Semantic Web is almost twenty years old (Berners-Lee, Hendler, and Lassila 2001). Since the inception of the World Wide Web, and to a great extent until today, its basic building block has been text documents connected by hyperlinks. These are documents that are intended for people, which need to be interpreted, and which may contain various pieces of information. The Semantic Web, by contrast, connects data instead of documents, for the use of computers. Data is here understood as a single piece of information that is machine readable. In contrast to documents, data cannot rely on human interpretation and therefore need to represent all meaning explicitly. The concept of Linked Data constitutes the mechanisms, technologies and frameworks for publishing data on the Semantic Web (Bizer, Heath, and Berners-Lee 2009). In RDF, each component of a triple can be a URI. Each piece of data, therefore, becomes globally addressable and reusable. Certain databases, such as relational databases, usually rely on their internal identifiers for database entries and use text labels to identify database fields. Using URIs instead means that both the entities in a database and what we say about those entities can be universally interpreted. Querying RDF data can be done through the SPARQL query language (W3C 2014). SPARQL queries can be stored as text files along with the RDF data for later reuse. Merely applying Linked Data principles is, however, no guarantee of reusable data. Linked Data is "not enough for scientists" (Bechhofer et al. 2013) without "a common model for describing the structure of our Research Objects including aspects such as lifecycle, ownership, versioning, etc." (Bechhofer et al. 2013, 569). Conceptual models such as the CIDOC-CRM described above are therefore crucial for creating truly reusable data, as are feasible methods for putting them into practice.

The challenge of maintaining digital humanities research outputs

Digital humanities produce a variety of digital artefacts that constitute the outcome of a research project: databases, websites, digital editions, virtual exhibitions, just to name a few. Such artefacts are often not only static files that can be stored but constitute pieces of software that must run and must be maintained for them to keep running as digital systems change and evolve. Technical debt is accumulated, as digital research outputs should remain reusable for future research (Hughes, Constantopoulos, and Dallas 2016, 161): "The use of ICT methods requires good practice in all stages of the digital life cycle to ensure effective use and reuse of data for research. Building digital collections of data for research involves consideration of the subsequent use and reuse of these collections for scholarship, using a variety of digital methods and tools".

Building a website constitutes not just a one-time effort but a long-term commitment, as Crymble (2015) observes: "Websites are expensive and a lot of work. Committing to building a website is like committing to build and maintain a library for the foreseeable future".

The disappearance of websites from the internet is a common phenomenon. Sampath Kumar and Manoj Kumar (2012) reviewed the decay of online citations in open access journals and found that almost a third of the cited articles were no longer accessible. This is a serious problem when those websites constitute valuable research outputs that are often also the result of significant financial investment. When funding stops, websites disappear (Bicho and Gomes 2016): "Most current Research & Development (R&D) projects rely on their websites to publish valuable information about their activities and achievements. However, these sites quickly vanish after the project funding ends".

The commitment required in maintaining online publication and research outputs is often overlooked in digital humanities scholarship (Reed 2014): "[…] coursework and publications related to DH project management tend to focus heavily on the difficulties of planning and launching a new project rather than the challenges of maintaining an established one […]"

As Bethany Nowviskie (2012) writes, digital humanities tend to emanate a feeling of "Eternal September", referring to the September influx of new students where all is new and fresh and everything can be built from scratch. This feeling ignores the fact that the digital artefacts that we build require maintenance and conveniently overlooks everything that is already available and needs taking care of. The notion of an "Eternal September", as Ashley Reed writes, "can also give the mistaken impression that digital humanities projects are inherently disposable: that long-term project management is unnecessary because creating a project is more important than developing or sustaining it" (Reed 2014, para. 2).

As the Web became commonplace, digital humanities researchers started to use more "sophisticated" tools. With this, the likelihood that digital research artefacts would become defunct increased significantly, either because they depend on the operation of underlying infrastructures such as databases and web servers (e.g., in content management systems such as WordPress or Drupal) or because the technology itself became obsolete (e.g., Flash). Sperberg-McQueen and Dubin (2017) describe a layered dependence of research artefacts on digital infrastructure: "In existing computer systems there is typically a long chain of relations connecting the physical phenomena by which data are represented with the data being represented. Each link in the chain connects two layers of representation: each layer organizes information available at the next lower level into structures at a higher (or at least different) layer of abstraction, and in this way provides information used in turn by the next higher level in the representation".

The layers ascend from the physical representation of data on storage devices to application-specific data structures and then to the presentation layer. With increasing numbers of layers, the long-term availability of digital research outputs becomes more difficult, as each layer depends on the previous ones and requires dedicated maintenance. It is due to such observations that a shift towards "the application of minimalist principles to computing" under the framework of minimal computing (Go::DH 2014; Varner 2017) is being promoted by some scholars (Gil and Ortega 2016). Minimal computing refers to "computing done under some set of significant constraints of hardware, software, education, network capacity, power, or other factors" (Go::DH 2014). In practice, this can mean publishing a website not through a database server that requires constant maintenance, upkeep and an internet connection but instead as a set of static documents that could be distributed on a USB stick in communities where internet access is scarce. Using the analogy of Sperberg-McQueen and Dubin (2017), minimal computing aims to reduce the layers of data representation that must be maintained, thereby making the challenge of producing digital research outputs that remain usable in the long term more realistic to achieve.

The challenge of long-term preservation

When facing the challenges of long-term preservation of research data, it is worth taking a closer look at the GAMS infrastructure of the University of Graz (Stigler et al. 2018). Their situation is in some ways comparable to the one we face at our institute and which we discuss below. After maintaining a proprietary pool of research-supporting software projects and technology, which became more costly and difficult over time, all projects then existing were transferred to a single environment for long-term archiving and provision of scientific data and content. The goal is to ensure sustainable availability and flexible (re-)use of digitally annotated and enriched scientific content. This is achieved through a largely XML-based content strategy based on domain-specific data models. Separation of the content and its presentation is an integral part of the infrastructure's architecture. Using recognised international standards like TEI, LIDO, SKOS, EDM or Dublin Core, Stigler et al. (2018) emphasise in their paper that the challenges of long-term preservation of research data cannot be solved without strong commitment from academic institutions, which have to perceive them as their central responsibility.

In her paper Research Data Management Instruction for Digital Humanities, Dressel (2017) states that, despite the interest in data curation in the digital humanities, little attention has been paid to providing instruction in research data management for the digital humanities: "Data curation represents a full range of actions on a digital object over its lifecycle and includes the basics of data management" (Dressel 2017, 8). To achieve successful long-term research data preservation, she emphasises the importance of the strong collaboration between librarians, researchers and IT staff.

In their book Cinderella's Stick: A Fairy Tale for Digital Preservation, Tzitzikas et al. (2018) point out the great importance of digital preservation, describing at the same time the challenges that come with it, which are different for all different types of digital artefacts. "While on the one hand, we want to maintain digital information intact as it was created; on the other, we want to access this information dynamically and with the most advanced tools" (Tzitzikas et al. 2018, 2). With regard to research data, they claim the usage of semantic web technologies such as RDF and the existing ontologies are beneficial: "Overall, we could say that the Semantic Web technologies are beneficial for digital preservation since the 'connectivity' of data is useful in making the semantics of the data explicit and clear. This is the key point for the Linked Open Data initiative, which is a method for publishing structured content that enables connecting it" (Tzitzikas et al. 2018, 65).

Case study: Self-contained research data at scale

At the Max Planck Institute for the History of Science (hereafter the Institute) we know all too well the amount of effort involved in maintaining digital research outputs (as well as the consequences of not being able to do so). The Islamic Scientific Manuscripts Initiative (ISMI, Daston et al. 2006), e.g., is one of the longest-running digital projects at the Institute. It constitutes a database catalogue of Islamic scientific manuscripts, including digitised sources where available (Daston et al. 2006). At its conception in 2006, a custom database was developed because no existing solution permitted the representation of the manuscripts and their scholarly and social connections at the level of detail that the scholars required. Currently, the data is being migrated into a CIDOC-CRM data model stored in an RDF triple store, as these models and technologies have matured and become widely available (Kuczera 2018). The ability to keep this unique source accessible has, however, hinged on both the availability and funding of a dedicated IT specialist throughout the project's lifetime up until today.

This has not been possible for the large majority of digital projects that have been developed, e.g. in collaboration with visiting researchers who have since left the Institute. An internal survey has unearthed 125 digital projects (and counting) residing on the Institute's servers. While many of them are surprisingly still operational – largely in cases where they have been built as static HTML websites – this is neither to be taken for granted nor relied on.

About a fifth of the 125 projects we identified at our own Institute have by now been either retired or stabilised, some of them as isolated run-time environments in which a project's state is conserved while the security risk of running outdated software is mitigated. This solution is acceptable if we want to preserve a website as an outcome of a research project. It is insufficient, however, when our goal is to allow future researchers to build on and reuse the digital artefacts that have been created.

When we regard digital projects as research outputs that should be shared and reused, our focus is therefore not on the presentation layer in the form of a website or an interactive data visualisation; instead, it is on the data that have been collected and curated to realise the presentation, i.e., the research data. This shift from presentation-centred digital scholarship towards an awareness of the value of data-driven approaches can be seen in recent calls for thinking of digital collections as collections of data and for a rethinking of cultural institutions as data-brokers (Ziegler 2020).

In shifting from presentation to data-centred digital humanities projects, though, we encounter two main problems for preservation. Firstly, many projects do not delineate a presentation layer from a data layer. This has been the case especially in projects employing technologies such as Adobe/ Macromedia Flash, which allow a project to be published as a single file. The situation slightly improved with the adoption of content management systems, where data is stored in a database. But the underlying database generally remains inaccessible to users who are only able to access it through predefined views or search interfaces. The second problem is that not everything that constitutes research data is expressed in data. A more recent digital project completed at the MPIWG is Sound & Science: Digital Histories (Tkaczyk et al. 2018). This website collects digital sources related to the history of acoustics and presents them through a search interface and in thematic sets, while also contextualising them in many written essays. It is based on the content management system Drupal and everything is stored in a database. Nevertheless, it is only through the presentation layer, where objects, images and texts are drawn together through customised views and database queries, that meaningful contexts are established. In a relational database model such as the one on which Drupal relies, individual entities are stored in separate tables. For example, the database entry describing a particular source and the database entry describing the person who authored that source reside in two different tables. And while these entities are linked together through an identifier on the database level, it is only through a database query and subsequent visual presentation that, e.g., the meaning of a relationship between a person and a text as that of "authorship" becomes evident to the user.

This is a central problem that we identified in several digital humanities projects, both of our own making and within the field: the full value of a digital research output manifests itself only through the combination of data and business logic (in the form of database queries and custom views). Research outputs rely on several layers of abstraction, as we outlined above with reference to Sperberg-McQueen and Dubin (2017). The upper layers provide meaning to the former. How, then, can we create research data that can live on its own, separate from software interfaces that might provide context to human users, but that we are unable to maintain?

From a library perspective, the transition to new digital publication environments and (micro)formats for publication as described above have changed the traditional workflows of collecting, cataloguing and archiving research outputs. Libraries have reliably accumulated publications over centuries and thus secured the functioning of the research life cycle. The life cycle is based on scholarly publications, which build on existing publications and flow back into retrieval and archival systems. This previously well-functioning life cycle of creating, publishing, evaluating, disseminating, archiving and retrieving has long since cracked: research data, the content of databases, websites and data visualisations do not flow back reliably into retrieval mechanisms anymore and are at risk of vanishing.

In our case study, we present two projects that have been developed in parallel for four years beginning in 2016: the Max Planck Digital Research Infrastructure and the research project The Sphere: Knowledge System Evolution and the Shared Scientific Identity of Europe. The goal of the Max Planck Digital Research Infrastructure is to establish conceptual workflows and technical infrastructure for storing semantically rich research data, linking it with relevant digital sources and providing user interfaces and APIs to keep data usable even after a project has ended. Within the Sphere research project, we tested conceptual and technical approaches for creating self-contained Linked Data according to the CIDOC-CRM standard, which in turn informed the design of the Max Planck Digital Research Infrastructure, the second project we present in this case study.

The Sphere: Knowledge system evolution and the shared scientific identity in Europe

The Sphere project revolves around the history of a single text: the *Tractatus de Sphaera* written by Johannes de Sacrobosco (Valleriani 2017, 2020). Sacrobosco's *Tractatus* is a short treatise on geocentric cosmology written during the 13th century, which gave rise to a very successful commentary tradition. It was usually published together with other texts taken from different disciplines that were seen as relevant for the study of cosmology. Within the project to date we have collected digital copies of 356 editions in which this particular text appears. The corpus begins with the earliest printed edition published in 1472 and spans a timeframe of roughly 180 years until the mid-17th century when the relevance of the *Tractatus* rapidly declined. What the project seeks to investigate based on this corpus is how certain texts and the knowledge that they conveyed have been disseminated, and what the contributing factors were that supported or hindered the spread of certain kinds of knowledge. The project has resulted in new findings on epistemic communities within the corpus (Valleriani et al. 2019; Zamani et al. 2020).

To identify the possible influence of certain factors such as individual publishers, the composition of each book, the location of printers or the language in which an edition was published, we need to store relevant data in a way that allows us to identify and trace arbitrary connections between them. This is the issue that we outlined in the first part of our chapter: meaning is found not in the individual entities but through how relationships are established between them. While the project began collecting bibliographic data about the corpus in a relational database, it was clear that a change of architecture would be required and that semantically Linked Data would be a crucial element for realising this research project.

Using the CIDOC-CRM ontology and the FRBRoo extension for bibliographic records (Bekiari et al. 2015), we created an initial data model for representing the bibliographic records of our corpus. Following the FRBR paradigm (Madison et al. 1997), an individual book is represented as separate components representing the physical copy (item), the printing template (manifestation) and the included text (expression). Using RDF, we can represent each component, as well as the events and actors that are associated with them, as individually addressable entities. Treating the content of a book as an entity on its own, which can, in turn, include other entities, allowed us to model a detailed representation of the individual texts that each edition contains. We could adapt and extend the data model as our understanding of the corpus grew and as new research questions arose. For instance, we could identify when individual texts were derived from other texts through processes of annotation or translation, thereby modelling entire genealogies of texts.

Representing the corpus in semantically rich RDF allows for a self-contained dataset, so that meaning is encoded in the data itself and not only at the point of retrieval via appropriate queries and presentation through suitable user interfaces. However, the meaning that is no longer being extracted at data output, therefore needs to be made manifest at data input, increasing the complexity of data entry. For the Sphere project, we built a data entry platform based on the Metaphacts system (Metaphacts 2019). This platform supports form-based data entry and image annotation as well as query and visualisation tools. From the perspective of a researcher, the platform's interface does not therefore significantly differ from common database-entry forms, preventing researchers from having to interact with the RDF data directly. A public instance of the platform can be found online and is documented in Kräutli and Valleriani (2018).

While the platform features a visual interface for composing custom queries, we found the availability to query RDF data directly via SPARQL to be the most useful for our research. We can query the data from Jupyter Notebooks (Project Jupyter 2020). Jupyter Notebooks are a text-based file format in which code can be combined with textual explanations, creating executable notebooks or even scholarly articles with embedded executable code. In the Sphere project, we employed Jupyter Notebooks to combine data query, analysis and visualisation in a shareable and self-contained format. Once we can no longer maintain our data entry platform, it will still be possible to download a copy of the project's research data. The notebooks can still be used locally to analyse the data. Instead of creating software that needs to be maintained and hosted, we create static artefacts that can be stored.

In the Sphere project, we applied several ideas that have been suggested for addressing the challenges of data reusability and preservation, such as Linked Data and the CIDOC-CRM data model. Implementing practical realisations of these paradigms gave us valuable insights into how we can address the issue of research data preservation at scale and create workable solutions for maintaining access to digital research outputs. Designing those solutions was the goal of the Max Planck Digital Research Infrastructure Project.

The Max Planck Digital Research Infrastructure

The ambition of the Max Planck Digital Research Infrastructure is to complete the digital research life cycle and to address the problems outlined above. We therefore designed an infrastructure to address an immediate need: the ability to maintain digital humanities research outputs so that they remain accessible and usable in the long term.

The most crucial realisation for achieving this is that most of what has been created at the Institute for digital humanities projects, and what we called databases, websites or visualisations at the time, is in fact software. Software needs to run, needs to be kept running and therefore needs constant maintenance. Lacking the resources for this, we need to separate data from software, creating self-contained datasets as demonstrated in the Sphere project. The painful consequence of this reality is that most of the user interfaces we create, most of the interactive visualisations that provide engaging access to research outputs, will not be around forever. Creating digital research outputs that remain usable also means designing the end of life of many artefacts that we create.

Our infrastructure comprises four main components: a repository, working environments, a data archive and a knowledge graph. The repository is a store of digitised sources, the Institute's digital collection. Scholars conduct their research within working environments that contain project-specific tools and artefacts. While researchers are working on a project, they use specific software and custom interfaces that may not be usable and maintainable in the long term and that will therefore be switched off at the end of a project. What remains after a project has ended is the research data, which is stored in the data archive. From there, it is fed into an institute-wide knowledge graph, where it is combined with sources in the repository as well as with data from previous research projects.

The knowledge graph becomes a central access point for all our digital artefacts, be they digitised sources, annotations or the datasets created within research projects. For this heterogeneous data to be compatible with other data, they need to be aligned to a common data model. This is where Linked Data principles come into play, namely the use of unique identifiers (URIs) to represent the same objects, together with the CIDOC-CRM ontology. We have successfully employed these principles in previous research projects. However, applying them at scale to all our research data is a new challenge that we have yet to face. Aside from the technical hurdles, working with these models requires a new set of skills in data modelling and knowledge representation, for which librarians and digital humanities developers need to be prepared.

Discussion: Implications and future challenges

What are the lessons learned in these projects and what are the next steps? Which changes and developments do we envisage in the field of digital humanities and digital data curation in the future?

What we have found is that we are not alone in the challenges that we face. The issues and unsolved questions surrounding data legacy that we are struggling with are the same as those confronting many research and cultural heritage institutions worldwide. In every presentation we gave in recent years, we received a great deal of feedback along with many questions and requests for further exchange of expertise. It seems that most of these institutions have reached a point where the number of legacy projects has become so significant, and the danger of vanishing data so pressing, that the search for a solution has become a considerable priority and the appetite for change, along with its disruptive implications, is increasing. This also explains why a growing community is evolving around Linked Data front ends, as described above. Linked Data, as we were able to show in the previous section, is certainly a suitable solution that can separate research data from software, interlink research data with the sources to which it refers, and, most importantly, let the data flow back into the digital research life cycle. Yet we must also acknowledge that there remain problems to be solved.

In addition to the technical challenges, we faced organisational stumbling blocks when we sought to follow an agile development approach within the administrative framework of a public institution. Since we started the digital research infrastructure project with a full set of open questions that needed to be solved along the way, an agile approach was unavoidable. Unfortunately, the administrative guidelines of public institutions are not entirely compatible with agile approaches, in the case of Germany, at least. These require the project to specify the exact software requirements and individual stages of development in detail before a contract with any company can be made. It took us some weeks to do this and many workarounds with colleagues from the Institute's administration needed to be sorted out in order to make our agile approach possible.

Another challenge is the undoubtedly steep learning curve that all project members face in gaining practice with data modelling using CIDOC-CRM or other compatible models such as FRBRoo. While Linked Data has been widely adopted by libraries over the last decade to describe their bibliographic data and interlink it with authority data, our projects aim to model not only the bibliographic records in a Linked Data format but the research data as well. This contributes to sustainably securing the data and – in the longer term – to being able to turn the "web of documents" into a "web of data".

Significant and long-term commitment and investment from all participants is therefore crucial for the successful outcome of these kinds of projects: librarians have to build up expertise in data modelling, ontologies and domain-specific vocabularies and take the lead in these fields within the research projects. Especially in the humanities, librarians must adjust more and more to new paradigms of research outcomes such as research data and other micro-publications. In order to address the question of how collaborations between information studies and digital humanities will progress and deepen, we argue that a broader knowledge of data modelling and existing ontologies will need to form a part of the curriculum for information studies. Librarians have always been experts in metadata; becoming experts in data modelling is nothing but the necessary next step. The significant difference is that their work must now become part of the research process at a much earlier stage and not only once the work is finished. Only as part of the project team can librarians advise scholars on how to express their research data using controlled vocabularies and ontologies, while also showing them the benefits of doing so. This approach will enable and require new ways of collaborating and stronger interactions between library professionals, IT experts and researchers. It can be said in general that the growth of digital approaches in the humanities is inevitably leading to more teamwork since different kinds of expertise are needed. In our experience, all sides benefit immensely from this collaboration.

While it will be the responsibility of the librarians to provide guidance, to maintain library data in ways that interconnect with research projects, and to establish standard interfaces for the exchange of research data, humanities scholars have to face the challenge of developing their projects within a digital framework and exploring digital methods from the very beginning. Using Linked Data paradigms at an early stage opens up many opportunities to exploit the data later. This will have a profound impact on research processes and methodologies. Following this approach represents a step towards genuine digital research in the humanities: digital research that rightly deserves the name "digital humanities". These approaches also need to be reflected in curricula within the humanities.

Last but not least, this deep collaboration between research projects and research infrastructures will lead to a shift of responsibilities, especially between the institution's academic staff in the digital humanities, IT specialists and the library. In our case, we are still in the process of (re)defining workflows, tasks and duties between the units as the projects evolve further. A central idea within this is to roll out every DH project in small teams consisting of the researchers, IT research staff and a librarian to provide support for data modelling.

When we started our infrastructure project, it was mainly driven by questions of maintenance and sustainability from the perspectives of the library and of IT staff: these are very practical questions about library infrastructures and about how to avoid falling into the same traps again, migrating and securing research projects and their data and finding a solution to make solving these challenges easier in future. We discovered early on that the data should be produced in certain formats, following certain data policies and using suitable ontologies. That required much closer interaction with the researchers and the research process itself than we initially expected. What seemed at first to be a sort of by-product became the centre of our attention and of enormous benefit to all parties involved: the close collaboration with the research project had a significant impact on the research methods that were employed, which enabled the researchers to work in a genuinely digital manner from the very beginning of their project. This led us to another important insight: in aiming for stable workflows and infrastructures, for clear structures and responsibilities, we had to accept the fact that the whole field that we work in is constantly developing and changing. Its fluidity is not only a result of the fact that each of the different disciplines engaged in the process (humanities, information science and IT) is very much in transition in terms of DH tools and methods, but also because the nature and degree of collaboration required in this framework are new to all three of them within this paradigm. Developing our project further than this is something we have to take into consideration. A certain level of flexibility is required not only throughout the project but also in more general terms, since we cannot initially predict all of the demands that will be placed upon our infrastructure and workflows. Our pilot project The Sphere provides a clear example: having been built on Linked Data principles and using CIDOC-CRM from the start, it provided an excellent use case for our digital research infrastructure. As the project evolved, it developed in a highly innovative direction, using methods of machine learning, identifying certain clear patterns throughout the history of the printing of the *Tractatus de Sphaera*. Using our data framework as a basis, the project went in a direction that we could not have predicted. This constant openness to changing paradigms is undoubtedly a challenge for research units and their infrastructures. At the same time, a constant dialogue between all participants is required, namely between humanities researchers and the supporting infrastructures for their research. We certainly intend to continue developing these in the future and, in our view, herein lies enormous potential for the development of digital tools and methods in the humanities.

Bibliography

Anderson, Sheila. 2013. “What are Research Infrastructures?” *International Journal of Humanities and Arts Computing* 7 (1–2): 4–23.

Bechhofer, Sean et al. 2013. “Why Linked Data is not Enough for Scientists.” *Future Generation Computer Systems* 29 (2): 566–611.

Bekiari, Chryssoula, Martin Doerr, Patrick Le Boeuf and Pat Riva, eds. 2015. *Definition of FRBRoo: A Conceptual Model for Bibliographic Information in Object-Oriented Formalism*. International Federation of Library Associations and Institutions (IFLA). https://www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Semantic Web." *Scientific American* 284 (5): 34–43.

Bicho, Daniel, and Daniel Gomes. 2016. "Preserving Websites of Research & Development Projects." IPRES.

Bizer, Christian, Tom Heath, and Tim Berners-Lee. 2009. "Linked Data: The Story So Far." *International Journal on Semantic Web and Information Systems* 5 (3): 1–22.

Borgman, Christine L. 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Champion, Erik Malcolm. 2014. "Researchers as Infrastructure." In *Proceedings of the Digital Humanities Congress 2012: Studies in the Digital Humanities*, edited by Clare Mills, Michael Pidd, and Esther Ward. Sheffield: The Digital Humanities Institute. https://www.dhi.ac.uk/openbook/chapter/dhc2012-champion

Champion, Erik Malcolm. 2018. "Introduction: A Critique of Digital Practices and Research Infrastructures." In *Cultural Heritage Infrastructures in Digital Humanities*, edited by Agiatis Benardou, Erik Champion, Costis Dallas, and Lorna M Hughes, 1–14. London: Routledge.

Crofts, Nick, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff, eds. 2011. *Definition of the CIDOC Conceptual Reference Model*. Heraklion: ICOM/CIDOC CRM Special Interest Group.

Crymble, Adam. 2015. "Does Your Historical Collection Need a Database-Driven Website?." *Digital Humanities Quarterly* 9 (1). http://www.digitalhumanities.org/dhq/vol/9/1/000206/000206.html

Daston, Lorraine, Jamil Ragep, Sally Ragep, and Robert Casties. 2006. "Islamic Scientific Manuscript Initiative." https://ismi.mpiwg-berlin.mpg.de/

Dierig, Sven, Jörg Kantel, and Henning Schmidgen. 2000. "The Virtual Laboratory for Physiology." Preprint 140. Max Planck Institute for the History of Science.

Doerr, Martin, and Nicholas Crofts. 1999. "Electronic Esperanto: The Role of the Object Oriented CIDOC Reference Model." In Proceedings of the ICHIM '99, Washington DC.

Doerr, Martin, Stefan Gradmann, Patrick LeBoeuf, Trond Aalberg, Rodolphe Bailly, and Marlies Olensky. 2013. *Final Report on EDM - FRBRoo Application Profile Task Force*. Europeana. http://pro.europeana.eu/taskforce/edm-frbroo-application-profile.

Doerr, Martin, Stephen Stead, and Maria Theodoridou. 2016. Definition of the CRMdig. FORTH.

Dressel, Willow. 2017. "Research Data Management Instruction for Digital Humanities." *Journal of EScience Librarianship* 6 (December): e1115. https://doi.org/10.7191/jeslib.2017.1115

Gil, Alex, and Élika Ortega. 2016. "Global Outlooks in Digital Humanities: Multilingual Practices and Minimal Computing." In *Doing Digital Humanities*, edited by Constance Crompton, Richard J. Lane, and Ray Siemens. London: Routledge.

Go::DH. 2014. "Minimal Computing." 2014. https://go-dh.github.io/mincomp/

Goerz, Guenther, Martin Scholz, Dorian Merz, Siegfried Krause, Mark Fichter, Kerstin Reinfandt, Peter Grobe, and Maria AnnaPfeifer. 2009. "What is WissKI?" http://wiss-ki.eu/what_is_wisski

Henry, Geneva. 2014. "Data Curation for the Humanities." In *Research Data Management: Practical Strategies for Information Professionals*, edited by Joyce M. Ray, 347–374. DGO-Digital original. West Lafayett: Purdue University Press. https://doi.org/10.2307/j.ctt6wq34t.20

Hughes, Lorna, Panos Constantopoulos, and Costis Dallas. 2016. "Digital Methods in the Humanities: Understanding and Describing Their Use across the Disciplines." In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 150–170. Hoboken, NJ: John Wiley & Sons.

IFLA. 1998. *Functional Requirements for Bibliographic Records*. Munich: K. G. Saur Verlag. https://www.ifla.org/publications/functional-requirements-for-bibliographic-records

Kilchenmann, Andre, Flavie Laurens, and Lukas Rosenthaler. 2019. "Digitizing, Archiving… and Then? Ideas about the Usability of a Digital Archive." In *Archiving Conference, Archiving 2019 Final Program and Proceedings* (5): 146–150. Society for Imaging Science and Technology. https://doi.org/10.2352/issn.2168-3204.2019.1.0.34

Kräutli, Florian, and Matteo Valleriani. 2018. "CorpusTracer: A CIDOC Database for Tracing Knowledge Networks." *Digital Scholarship in the Humanities* 33 (2): 336–346. https://doi.org/10.1093/llc/fqx047

Kuczera, Andreas. 2018. *Graphentechnologien in den digitalen Geisteswissenschaften*. GitHub. https://kuczera.github.io/Graphentechnologien/

LOC. 2002. "EAD: Encoded Archival Description (EAD Official Site, Library of Congress)." https://www.loc.gov/ead/

LOC. 2016. "Overview of the BIBFRAME 2.0 Model (BIBFRAME – Bibliographic Framework Initiative, Library of Congress)."
https://www.loc.gov/bibframe/docs/bibframe2-model.html

Madison, Olivia, John Byrum, Suzanne Jouguelet, Dorothy McGarry, Nancy Williamson, Maria Witt, Tom Delsey, Elizabeth Dulabahn, Elaine Svenonius, and Barbara Tillett. 1997. "Functional Requirements for Bibliographic Records." *International Federation of Library Associations and Institutions*. https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf

MARBI. 1996. "The MARC 21 Formats: Background and Principles." https://www.loc.gov/marc/96principl.html

Metaphacts. 2019. "Metaphacts-Community." https://bitbucket.org/metaphacts/metaphacts community

Nowviskie, Bethany P. 2012. "Eternal September of the Digital Humanities." In *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press. https://dhdebates.gc.cuny.edu/read/untitled-88c11800-9446-469b-a3be3fdb36bfbd1e/section/4aaed3b0-07ed-4c4e-ad49-8f7ceecf740a.

O'Sullivan, James. 2019. "The Humanities Have a "Reproducibility" Problem." *Talking Humanities*. 9 July 2019. https://talkinghumanities.blogs.sas.ac.uk/2019/07/09/the-humanities-have-a-reproducibility-problem/

Oldman, Dominic. 2016. "ResearchSpace." https://public.researchspace.org/resource/Start

Oldman, Dominic, Martin de Doerr, Gerald de Jong, Barry Norton, and Thomas Wikman. 2014. "Realizing Lessons of the Last 20 Years: A Manifesto for Data Provisioning and Aggregation Services for the Digital Humanities (A Position Paper)." *D-Lib Magazine* 20 (7/8). https://doi.org/10.1045/july2014-oldman

Oldman, Dominic, Martin Doerr, and Stefan Gradmann. 2016. "Zen and the Art of Linked Data: New Strategies for a Semantic Web of Humanist Knowledge." In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens and John Unsworth, 251–273. Hoboken, NJ: John Wiley & Sons.

Peels, Rik and Lex Bouter. 2018. "The Possibility and Desirability of Replication in the Humanities." *Palgrave Communications* 4 (1): 95. https://doi.org/10.1057/s41599-018-0149-x

PARTHENOS. n.d. "The Data Heterogeneity Problem – Parthenos Training." Accessed 23 March 2021. https://training.parthenos-project.eu/sample-page/formal-ontologies-a-complete-novices-guide/what-is-data-heterogeneity/.

Project Jupyter. 2020. "Project Jupyter." https://www.jupyter.org.

Reed, Ashley. 2014. "Managing an Established Digital Humanities Project: Principles and Practices from the Twentieth Year of the William Blake Archive." *Digital Humanities Quarterly* 8 (1). http://www.digitalhumanities.org/dhq/vol/8/1/000174/000174.html

Renn, Jürgen. 2002. "Echo: Ein Virtueller Marktplatz für das Kulturelle Erbe." *Max-Planck-Forschung*, (2): 68–74.

Sampath Kumar, B. T., and K. S. Manoj Kumar. 2012. "Decay and Half-Life Period of Online Citations Cited in Open Access Journals." *International Information & Library Review* 44 (4): 202–11. https://doi.org/10.1080/10572317.2012.10762933

Sperberg-McQueen, C. M., and Dubin, David. 2017. "Data Representation." In *Digital Humanities Data Curation*. Accessed 19 March 2021. https://guide.dhcuration.org/contents/data-representation/.

Stigler, Johannes Hubert, and Elisabeth Steiner. 2018. "GAMS–An Infrastructure for the Long-Term Preservation and Publication of Research Data from the Humanities." *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare* 71 (1): 207–16.

Tkaczyk, Viktoria, Joeri Bruyninckx, Fanny Gribenski, Xiaochang Li, Kate Sturge, Robert Casties, and FlorianKräutli. 2018. "Sound & Science: Digital Histories." https://soundandscience.de/

Tzitzikas, Yannis and Yannis Marketakis. 2018. *Cinderella's Stick: A Fairy Tale for Digital Preservation*. Cham: Springer. https://doi.org/10.1007/978-3-319-98488-9

Valleriani, Matteo. 2017. "The Tracts on the Sphere: Knowledge Restructured over a Network." In *The Structures of Practical Knowledge*, edited by Matteo Valleriani, 421–473. Cham: Springer. https://doi.org/10.1007/978-3-319-45671-3_16

Valleriani, Matteo, Florian Kräutli, Maryam Zamani, Alejandro Tejedor, Christoph Sander, Malte Vogl, Sabine Bertram, Gesa Funke, and Holger Kantz. 2019. "The Emergence of Epistemic Communities in the Sphaera Corpus." *Journal of Historical Network Research* 3 (November): 50–91. https://doi.org/10.25517/jhnr.v3i1.63

Valleriani, Matteo, ed. 2020. *De Sphaera of Johannes de Sacrobosco in the Early Modern Period: The Authors of the Commentaries*. Cham: Springer. https://doi.org/10.1007/978-3-030-30833-9

Varner, Stewart. 2017. "Minimal Computing in Libraries: Introduction." https://godh.github.io/mincomp/thoughts/2017/01/15/mincomp-libraries-intro/

W3C. 2014. “RDF – Semantic Web Standards.” https://www.w3.org/RDF/

Zamani, Maryam, Alejandro Tejedor, Malte Vogl, Florian Kräutli, Matteo Valleriani, and Holger Kantz. 2020. “Evolution and Transformation of Early Modern Cosmological Knowledge: A Network Study.” *Scientific Reports* 10: 19822. https://doi.org/10.1038/s41598-020-76916-3

Ziegler, S. L. 2020. “Open Data in Cultural Heritage Institutions: Can We Be Better Than Data Brokers?” *Digital Humanities Quarterly* 14 (2). http://www.digitalhumanities.org/dhq/vol/14/2/000462/000462.html