Originally published in: Koschtial, Claudia/Köhler, Thomas/Felden, Carsten (Eds.): e-Science: Open, social and virtual technology for research collaboration – Cham: Springer, 2021. Pp. 67-76. (Progress in IS)
DOI: *https://doi.org/10.1007/978-3-030-66262-2_5*

Digital Research Infrastructure

Maik Stührenberg, Oliver Schonefeld, and Andreas Witt

Abstract Digital research infrastructures can be divided into four categories: large equipment, IT infrastructure, social infrastructure, and information infrastructure. Modern research institutions often employ both IT infrastructure and information infrastructure, such as databases or large-scale research data. In addition, information infrastructure depends to some extent on IT infrastructure. In this paper, we discuss the IT, information, and legal infrastructure issues that research institutions face.

Keywords Digital research infrastructure · IT infrastructure · Information infrastructure

1 Introduction

This paper was originally submitted late 2014 and the final publication was delayed until 2019. The authors are well aware that the view and state of the art for digital research infrastructures have evolved in the last 5 years.

A research infrastructure can be defined as a public or private institution that has been established mainly for research, teaching, and the support of young researchers. Research infrastructures can be divided into four main categories (Wissenschaftsrat 2011b, 17f.)[1]:

– large equipment, including research platforms such as scientific research vessels, planes, or satellites;

[1] Combinations of more than one category are possible as well.

M. Stührenberg (✉)
Universität Bielefeld, Bielefeld, Germany
e-mail: maik.stuehrenberg@uni-bielefeld.de

O. Schonefeld · A. Witt
Institut Für Deutsche Sprache, Mannheim, Germany
e-mail: schonefeld@ids-mannheim.de

A. Witt
e-mail: witt@ids-mannheim.de

C. Koschtial et al. (eds.), *e-Science*, Progress in IS,
https://doi.org/10.1007/978-3-030-66262-2_5

– IT infrastructure, such as computer hardware and software;
– social infrastructure, that is, research institutions that offer scholars a place to exchange ideas and collaborate (Wissenschaftsrat 2011a, 20f.), for example, the Leibniz Center in Dagstuhl Castle, Germany;
– information infrastructure, that is, research institutions that collect and curate primary data and make them accessible to a larger group of scholars.

While large technical equipment is only seldom used in digital humanities disciplines, and social infrastructure is beyond the scope of this paper, combinations of IT infrastructure and information infrastructure are quite common. Therefore, the purpose of this paper is to give insight into various aspects of modern research infrastructures with an emphasis on both the latter categories. In addition, we have conducted a qualitative analysis by interviewing twelve German research institutions (Fiedler et al. 2012). The institutions were interviewed and asked to participate in a survey. The 74 survey questions were structured into different topic areas, such as organizational aspects, data management, hardware and software, environmental aspects, and legal issues. We will reflect on some of these topics in the respective sections of this article.

2 IT Infrastructure

Digital humanities research institutions working with huge amounts of data (e.g., language corpora) have special needs regarding IT infrastructure, such as a growing demand for storage space, computing capacity (for querying and analyzing linked data), and durability (including distributed access over large-scale networks such as the Internet for a huge number of potential users). This results in significant amounts of money spent on hardware and software. In addition, operating costs (divided into maintenance and personnel costs) have to be taken into account, including IT staff, hardware maintenance, software updates, and licensing. Especially energy costs should not be underestimated, as the price of electricity is increasing over time. A green-IT strategy can help an institution to reduce some of these costs. A key way of doing this is buying new equipment and replacing old (less energy-efficient) hardware. However, green IT consists of more aspects, such as efficient cooling (like separation of warm and cold aisles in the data center or using free cooling techniques), institutional policies (e.g., obliging employees to turn their computers off before leaving the workplace), or using supplies made of recycled material (like recycled paper). Implementing a green-IT strategy is generally a project of its own for a research institution and is currently a low priority for the institutions that we analyzed.

Therefore, one of the issues modern-day research institutions have to deal with is to optimize these costs, usually by undertaking the following steps. Firstly, a transparent accounting system, including every single asset for salaries, maintenance costs, and so forth, has to be established, allowing for a more accurate estimation of current

and future demands for IT infrastructure. Replacing proprietary software with open-source software may only slightly decrease licensing costs, but may be cheaper in the long run since the latter can be adopted to the institution's needs and usually has better support of open formats (see Sect. 3.2). However, two points have to be considered regarding this assumption:

1. Additional costs for user training may be necessary if the open-source application differs from the formerly used product;
2. In-house IT expertise is necessary to adapt open-source software, which may result in even higher personnel costs.

For these reasons, it is advisable, especially for smaller research institutions, to collaborate in the field of IT infrastructure to reduce costs. Examples of such cooperation include a shared Internet connection, server housing, or archival storage. A majority of the interviewed research institutions already collaborate with other external facilities to lower IT costs and to distribute archival and backup storage.
Since research institutions are nowadays connected to the Internet, storage of and access to the information infrastructure involves special security requirements. Two main issues have to be considered:

1. preventing unauthorized access to systems, processes, or data (including information infrastructure);
2. ensuring that hardware and software continue to function.

Although there is no such thing as a completely secure network, the first step to prevent unauthorized access is a complete risk analysis for the relevant computer systems, including estimating possible losses and limitations on daily work (e.g., due to vandalism or sabotage). The outcome of this analysis should be a prioritized list of data and systems to be protected.

The concrete security measures (the security policy) are defined by the IT security officer and the data protection officer and are mandatory for the whole staff of the research institution (ISO/IEC 27002:2013 2013; BSI 2014). Important points for a security policy are:

– prioritization of data according to their value for the research institution;
– identification of possible risks (including computer viruses and network infrastructure attacks);
– backup strategy;
– data encryption.

While a backup strategy for research data is considered crucial (nine out of twelve interviewees have a central backup strategy and the remaining institution plans to implement one), only a third of the institutions surveyed have a central in-house IT security policy.

3 Information Infrastructure

Research data, especially primary data (e.g., recordings, measurements, and curated corpora), are among the most valuable assets for a research institution. Research institutions that can be categorized as information infrastructures (such as libraries, archives, collections, and smaller non-academic research institutions) that collect and curate primary data, scientific and non-scientific knowledge, and databases, and provide access to researchers [34], who may use this data for research projects on their own. To ensure access to the information infrastructure, various technical aspects have to be taken into account.

3.1 Repositories and Publication Server

Repositories have already been used in large-scale collaborative projects, often international ones, such as CLARIN.[2] The CLARIN centers provide repositories storing academic research data (such as curated corpora) accessible via the Internet. Retrieval of a desired information item is highly dependent on metadata. Following on from existing metadata standards such as Dublin Core (ISO 15836:2009 2009; DCMI 2012), IMDI (ISLE Metadata Initiative 2003; Broeder and Wittenburg 2006; ISLE Metadata Initiative 2009), or OLAC (Simons et al. 2008; Bird and Simons 2009), the Component Metadata Structure (CMDI) (Broeder et al. 2011, 2012; Trippel et al. 2012) has been created to facilitate documenting research information and querying it over the distributed repositories. In our survey, five out of the twelve interviewed institutes already run a repository on their own, while four are in the process of building one.

Another aspect of information infrastructure is the archiving and accessibility of publications. Establishing and maintaining an in-house publication server can be a way for a research institution to retain both copyright (see Sect. 4.1) and access control over information that has been produced by its academic staff. Open-source implementations, such as ePrints[3] or eSciDoc,[4] often combine the functionalities of publication servers and primary data repositories. For all these tasks, staff working on IT and information infrastructure need to collaborate closely. In particular, research institutions having their own libraries can benefit from the expertise of IT and information departments regarding archives, metadata, and retrieval. Seven of the interviewees already run a publication server.

[2] See http://www.clarin.eu for further details.

[3] See http://www.eprints.org/ for further details.

[4] See https://www.escidoc.org/ for further details.

3.2 Data Formats

Although the creation of research data is often quite expensive, a large portion of this information gets lost shortly after the end of the project in which it was gathered. Apart from the hardware failures or insufficient metadata discussed above, another possible reason can be a proprietary storage format, for which the corresponding application is not available any more.

Data formats usually exist for two reasons: (1) as serialization of a specification, or (2) as the import and export format of an application. A format as such may be open or proprietary, which may be important for processing and archiving the information encoded in it. An example of a proprietary de facto standard format is the ubiquitous.doc format, produced by Microsoft Word.[5] Since it is a binary format, it is not possible to extract information with arbitrary text editors; instead, one has to use specific programs, and applications other than MS Word may not be able to successfully render the document as it was intended by the author.

For research data which are curated by an information infrastructure, open text-based formats should be preferred. Formats based on the open meta language XML (Bray et al. 2008) are quite common in academic research and can be defined by document grammar formalisms such as XML DTD (part of the aforementioned specification), XML Schema (Gao et al. 2012; Peterson et al. 2012), or RELAX NG (ISO/IEC 19757-2:2008 2008), allowing for on-the-fly validation during the creation of instances. Examples of open XML-based annotation formats in the digital humanities are the TEI Guidelines (Burnard and Bauman 2014) or DocBook (Walsh 2010) for technical documentation. Information encoded in those formats is not only readable with common text editors, but separates content from formatting, since the rendering is usually controlled by separate XSLT (Kay 2007, 2014) or CSS (Bos et al. 2011) stylesheets. This not only prevents vendor lock-in, but significantly eases the process of archiving. The attitude to open standards and open-source software compared with proprietary in-house development is mixed; however, there is a tendency to use standardized APIs and formats, or at least consider open-source applications. Seven surveyed institutes keep data in proprietary formats, while four aim to use standard formats and one is still determining its strategy. Often, institutes lack the human resources to convert data into standard formats.

4 Legal Issues

Research institutions are confronted with a number of legal issues, the most important of which are: (1) copyright and (2) personal data protection and privacy.

[5]Note that we are talking about the binary .doc, not the XML-based .docx format used by Office 2004 onwards and that is standardized as ISO/IEC 29500-1:2011 (2011). However, even the latter format uses a number of features that cannot easily be interpreted by application programs without further knowledge.

4.1 Copyright Issues

Research data is often based on material contributed by third parties. The primary data of text corpora, for example, often originate from newspaper articles or similar non-academic sources. German copyright law protects literary, artistic, and scientific works (including software) that are the author's own intellectual creation. Copyright-protected works may only be modified (and, arguably, annotated) with the authorization of the copyright holder. Copyright expires 70 years after the death of the original author. In Germany (unlike in most other jurisdictions), copyright cannot be transferred and is reserved by the author until his death (and 70 years after it), but it can be licensed. In practice, authors often license their rights out to publishers.

Although the German copyright law (UhrG) does not contain the American concept of "fair use", there are copyright limitations (§§ 44a–63a UrhG) that apply to certain specific uses of copyright-protected works (e.g., citations, personal use, scientific use) (Mönch 2006). However, in order to be covered by a copyright limitation of § 52a UrhG, scientific use has to be restricted to "small groups of researchers" (Hoeren 2014, 157). This is especially important if a research institution wants to publish annotated corpora-in that case, the primary data has to be licensed beforehand.

Research data to which a research institution holds the copyright (e.g., primary data produced in-house) should be made available to others under a liberal license, e.g., an open-access license such as Creative Commons.[6] Creative Commons (CC) is a free license (similar to the software license, BSD,[7] or the General Public License, GNU[8]) that was originally developed for creative work and that consists of several building blocks, such as Attribution (BY: minimal requirement), NoDerivatives (ND), NonCommercial (NC),[9] and ShareAlike (SA). The current version (4.0) also addresses specific database rights that exist in EU Member States.

Apart from human-readable CC license deeds, laundry symbols (similar to those established in the CLARIN research group (Oksanen et al. 2010) for its own specific licenses) provide a quick overview of the license requirements.[10] For a detailed discussion about legal implications of institutional repositories see Bargheer et al. (2006).

Regarding publications, a research institution's staff may agree to publish their works on the institution's publication server under an open-access license (Degkwitz 2007). Open-access publications have steadily gained ground in countries such as the US, Denmark, or Japan, while there is still an ongoing discussion about them in Germany, especially in the digital humanities disciplines[11]—although the Berlin

[6] See http://creativecommons.org for further details.

[7] See http://opensource.org/licenses/bsd-license.php for further details.

[8] See http://www.gnu.org/licenses/#GPL for further details.

[9] Especially NC may have undesired side effects, see Klimpel (2012) for a discussion.

[10] The categories have recently been extended by Kupietz and Lüngen (2014).

[11] See Görl et al. (2011) for a discussion about the impacts of information infrastructure in universities of North Rhine-Westphalia.

Declaration on Open Access to Knowledge in the Sciences and Humanities[12] has boosted their reputation. While open-access journals are still sometimes seen as less reputable than traditional journals (although both publication types monitor quality through peer review), they often have higher citation numbers.[13] Research institutions can play an active role in the process of building the reputation of open access by publishing in this format. It is therefore pleasant to see that an open-access strategy is already present in five of the institutions interviewed, while three of them plan on implementing one.

4.2 Personal Data Protection

Personal data protection issues may arise when living persons are involved in the process of creating research data, such as voice or video recordings. Publication of personal data is only allowed if the persons recorded have given their (written) consent. For every collection of personal data, a register of processing operations has to be created (according to §4 g, §§18 and 4e of the German data protection law, BDSG. The type of personal information, how it is processed, and the data protection measures, are recorded in this register.

Despite the variety of legal issues that may arise for research institutions, most of the interviewees rely either on their own (general) legal department or on cooperation with external law firms. Licensed (IT law) attorneys are seldom employed. However, since German research institutions are required to employ a data protection officer if they deal with personal data, they already have at least some existing in-house expertise. This expert should be involved in any data collection activities as soon as possible.

5 Conclusion

We have discussed a number of information infrastructure issues that modern research institutions need to consider. Most of the technical issues can be addressed by implementing a sustainable long-term IT strategy that reflects both costs and demands. Additional technical aspects such as security, open storage formats, and metadata can be addressed in such an IT strategy. Legal issues cannot be underrated, especially for service-oriented research institutions. Therefore, a data protection officer should be involved in the early stages of research projects that plan to create personal data.

---

[12] See the text of the declaration at http://openaccess.mpg.de/3515/Berliner_Erklaerung.

[13] See Stempfhuber (2009, 119) and http://opcit.eprints.org/oacitation-biblio.html for a number of studies about open-access impact factors.

References

Bargheer, M., Bellem, S., Schmidt, B.: Open Access und Institutional Repositories– Rechtliche Rahmenbedingungen. In: Spindler, G. (ed.) Rechtliche Rahmenbedingungen von Open Access-Publikationen, pp. 1–20. No. 2 in Göttinger Schriften zur Internetforschung, Universitätsverlag Göttingen (2006)

Bird, S., Simons, G.F.: OLAC: Accessing the world's language resources. In: Proceedings of the 1st International Conference on Language Documentation and Conservation. Hawaii (2009)

Bos, B., Çelik, T., Hickson, I., Lie, H.W.: Cascading style sheets level 2 revision 1 (css 2.1) specification. W3C Recommendation, World Wide Web Consortium (W3C), http://www.w3.org/TR/2011/REC-CSS2-20110607 (2011)

Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., Yergeau, F.: Extensible Markup Language (XML) 1.0 (Fifth Edition). W3C Recommendation, World Wide Web Consortium (W3C), http://www.w3.org/TR/2008/REC-xml-20081126/ (2008)

Broeder, D., Schonefeld, O., Trippel, T., van Uytvanck, D., Witt, A.: A pragmatic approach to XML interoperability—the Component Metadata Infrastructure (CMDI). In: Proceedings of Balisage: The Markup Conference. Balisage Series on Markup Technologies, vol. 7. Montréal (2011)

Broeder, D., Windhouwer, M., van Uytvanck, D., Trippel, T., Goosen, T.: CMDI: a component metadata infrastructure. In: Arranz, V., Broeder, D., Gaiffe, B., Gavrili- dou, M., Monachini, M., Trippel, T. (eds.) Proceedings of the LREC 2012 Workshop on Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR, pp. 1–4. (2012)

Broeder, D., Wittenburg, P.: The IMDI metadata framework, its current application and future direction. Int. J. Metadata Semant. Ontol. **1**(2), 119–132 (2006)

Bundesamt für Sicherheit in der Informationstechnik: IT-Grundschutz- Vorgehensweise. Version 2.0. BSI-Standard 100–2, BSI, https://www.bsi.bund.de/SharedDocs/Doloads/DE/BSI/Publikationen/ITGrundschutzstandards/BSI-Standard_1002.pdf?__blob=publicationFile (2014)

Burnard, L., Bauman, S. (eds.): TEI P5: Guidelines for electronic text encoding and interchange. Text Encoding Initiative Consortium, Charlottesville, Virginia, version 2.6.0. Last updated on 20th January 2014, revision 12802 (2014)

DCMI Usage Board: Dublin Core Metadata Element Set, Version 1.1. DCMI Recommendation, Dublin Core Metadata Initiative, http://dublincore.org/documents/2012/06/14/dces/ (2012)

Degkwitz, A.: Open access und die Novellierung des deutschen Urheberrechts. Zeitschrift für Bibliothekswesen und Bibliographie **54**(4/5), 243–245 (2007)

Fiedler, N., Werthmann, A., Stührenberg, M., Bingel, J., Witt, A.: Forschungsinfrastrukturen in außeruniversitären Forschungseinrichtungen. Forschungsbericht, Institut für Deutsche Sprache. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-28325 (2012)

Gao, S.S., Sperberg-McQueen, C.M., Thompson, H.S.: W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures. W3C Recommendation, World Wide Web Consortium (W3C), http://www.w3.org/TR/2012/REC-xmlschema11-1-20120405/ (2012)

Görl, S., Puhl, J., Thaller, M.: Empfehlungen für die weitere Entwicklung der Wissenschaftlichen Informationsversorgung des Landes NRW. epubli (2011)

Hoeren, T.: Internetrecht. Universität Münster. https://www.itm.nrw/wp-content/uploads/Skript-Internetrecht-April-2014.pdf (2014)

ISLE Metadata Initiative: Metadata Elements for Session Descriptions. version 3.0.4. Reference Document, MPI, Nijmegen, https://tla.mpi.nl/wp-content/uploads/2012/06/IMDI_MetaData_3.0.4.pdf (2003)

ISLE Metadata Initiative: Metadata Elements for Catalogue Descriptions. version 3.0.13. Reference Document, MPI, Nijmegen, https://tla.mpi.nl/wp-content/uploads/2012/06/IMDI_Catalogue_3.0.0.pdf (2009)

ISO 15836:2009: The Dublin Core Metadata Element Set, Version 1.1. International Standard, International Organization for Standardization, Geneva (2009)

ISO/IEC 19757-2:2008: Information technology—Document Schema Definition Language (DSDL) – Part 2: Regular-grammar-based validation—RELAX NG. International Standard, International Organization for Standardization, Geneva (2008)
ISO/IEC 27002:2013: Information technology—Security techniques—Code of practice for information security controls. International Standard, International Organization for Standardization/International Electrotechnical Commission (2013)
ISO/IEC 29500-1:2011: Information technology—Document description and processing languages—Office Open XML File Formats—Part 1: Fundamentals and Markup Language Reference. International standard, International Organization for Standardization, Geneva (2011)
Kay, M.: XSL Transformations (XSLT) Version 2.0. W3C Recommendation, World Wide Web Consortium (W3C). http://www.w3.org/TR/2007/REC-xslt20-20070123/ (2007)
Kay, M.: XSL Transformations (XSLT) Version 3.0. W3C Last Call Working Draft, World Wide Web Consortium (W3C). http://www.w3.org/TR/2014/WD-xslt-30-20141002/ (2014)
Klimpel, P.: Freies Wissen dank Creative-Commons-Lizenzen. Folgen, Risiken und Nebenwirkungen der Bedingung 'nicht-kommerziell—NC'. Wikimedia Deutschland and iRights.info and Creative Commons Deutschland. http://irights.info/userfiles/CC-NC_Leitfaden_web.pdf (2012)
Kupietz, M., Lüngen, H.: Recent developments in DeReKo. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik (2014)
Mönch, M., Nödler, J.M.: Hochschulen und Urheberrecht – Schutz wissenschaftlicher Werke. In: Spindler, G. (ed.) Rechtliche Rahmenbedingungen von Open Access- Publikationen, pp. 21–54. No. 2 in Göttinger Schriften zur Internetforschung, Universitätsverlag Göttingen (2006)
Oksanen, V., Lindén, K., Westerlund, H.: Laundry symbols and license management—practical considerations for the distribution of LRs based on experiences from CLARIN. In: Arranz, V., van Eerten, L. (eds.) Language Resources: From Storyboard to Sustainability and LR Lifecycle Management, Workshop held at the seventh conference on International Language Resources and Evaluation (LREC 2010), pp. 10–13. Valletta (2010)
Peterson, D., Gao, S.S., Malhotra, A., Sperberg-McQueen, C.M., Thompson, H.S.: W3C XML Schema Definition Language (XSD) 1.1 Part 2: Datatypes. W3C Recommendation, World Wide Web Consortium (W3C). http://www.w3.org/TR/2012/REC-xmlschema11-2-20120405/ (2012)
Simons, G.F., Bird, S.: OLAC Metadata. Olac standard, Open Language Archives Community. http://www.language-archives.org/OLAC/metadata-20080531.html (2008)
Stempfhuber, M.: Die Rolle von "open access" im Rahmen des wissenschaftlichen Publizierens. In: Publikationsverhalten in unterschiedlichen wissenschaftlichen Disziplinen. Beiträge zur Beurteilung von Forschungsleistungen, pp. 116–131. Diskussionspapier der Alexander von Humboldt-Stiftung. 2 edn. Alexander von Humboldt Stiftung, Bonn (2009)
Trippel, T., Hoppermann, C., Depoorter, G.: The component metadata infrastructure (cmdi) in a project on sustainable linguistic resources. In: Arranz, V., Broeder, D., Gaiffe, B., Gavrilidou, M., Monachini, M., Trippel, T. (eds.) Proceedings of the LREC 2012 Workshop on Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR. pp. 29–36. (2012)
Walsh, N.: DocBook 5: The Definitive Guide. O'Reilly Media, Sebastopol (2010)
Wissenschaftsrat: Empfehlungen zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften. Berlin. http://www.wissenschaftsrat.de/download/archiv/10465-11.pdf (2011a)
Wissenschaftsrat: Übergreifende Empfehlungen zu Informationsinfrastrukturen. Berlin. http://www.wissenschaftsrat.de/download/archiv/10466-11.pdf (2011b)