INSTITUT FÜR
DEUTSCHE SPRACHE

Oliver Schonefeld **Maik Stührenberg** Andreas Witt

DIGITAL RESEARCH INFRASTRUCTURE

An overview

Leipzig

Member of the
Leibniz Association

STRUCTURE OF THIS TALK

1. Introduction
2. IT infrastructure
3. Information Infrastructure
   - Repositories and publication server
   - A question of formats
4. Legal issues
   - Copyright
   - Personal data protection
5. Conclusion

DIGITAL RESEARCH INFRASTRUCTURE

A Definition

A research infrastructure is the entirety of the facilities, information, resources and services that have been gathered for the sole purpose of research

(European Strategy Forum on Research Infrastructures 2006)

THE IDS AS DRI

- The Institute for the German Language was founded in 1964 and is located in Mannheim
- Member of the Leibniz society
- Curates the Archive of General Reference Corpora of Contemporary Written German (DeReKo) with currently more than 24 billion words (Kupietz and Lüngen 2014)
- Certified CLARIN centre
- Participating in nestor, TextGrid, DIN, and ISO, amongst others

CATEGORIES OF RESEARCH INFRASTRUCTURE

We can define four main categories of research infrastructure:

- Large research equipment, including scientific research vessels, planes or satellites
- IT infrastructure, such as computer hardware and software, Grids
- Social infrastructure, that is, research institutions, which offer scholars a place to exchange ideas and collaborate with each other
- Information infrastructure, that is, collections of research data that are made accessible for a larger group of scholars

CATEGORIES OF RESEARCH INFRASTRUCTURE

For the remainder of this talk only two of them are of interest for us:

1. IT infrastructure, and
2. Information infrastructure

Hint: The IDS fits in both categories

IT INFRASTRUCTURE

- Digital Humanities research institutions often work with huge amounts of data (e. g. language corpora)
- As a result they have special needs regarding IT infrastructure, such as
  - storage space
  - computing capacity (for querying and analyzing linked data)
  - durability (including distributed access over large-scale networks such as the Internet for a huge number of potential users)
- These special needs result in significant costs which can be differentiated into
  - Buying costs (hardware and software), and
  - Operating costs (maintenance, personell, energy)

COST CONTROL

Optimizing those costs can be done by …

- … establishing a transparent accounting system including every single asset for salaries, maintenance costs, and so forth
  → allows for a more accurate estimation of current and future IT infrastructure demands
- … replacing proprietary software with Open Source software
  → only small decrease in licensing costs, but may be cheaper in the long run because of better adaption to the institution's needs and better support of open formats

SECURITY CONCERNS

- Storage of and access to the information infrastructure requires special considerations regarding IT security
- Two main issues have to considered:
  1. Preventing non-authorized access to systems, processes or data (including information infrastructure)
  2. Providing a continued operation of hard and software
- To tackle these security concerns, security measures have to be taken into account

SECURITY MEASURES

- Concrete security measures (the security policy) are defined by the IT security officer and the data protection officer
- Important points of a security policy are:
  - Prioritization of data according to their value for the research institution
  - Identification of possible risks (including computer viruses, network infrastructure attacks, etc.)
  - Backup strategy
  - Data encryption
- Security policy is mandatory for the whole staff of the research institution (ISO/IEC 27002:2013)

TO SUM UP…

IT Infrastructure issues

To tackle the afore-mentioned IT Infrastructure issues regarding

- technical,
- financial, and
- security questions,

different groups of personnel of a research institution have to collaborate

1 Introduction

2 IT infrastructure

3 Information Infrastructure

- Repositories and publication server
- A question of formats

4 Legal issues

5 Conclusion

INFORMATION INFRASTRUCTURE

- Research data, especially primary data (e. g. recordings, measurements, and curated corpora), is one of the most valuable assets for a research institution
- To assure the access to the information infrastructure, various technical aspects have to be taken into account, including repositories, metadata, and formats

REPOSITORIES

- Repositories have already been used in large-scale collaborative projects (e. g. research groups such as CLARIN)
- In distributed projects, independent centres provide repositories storing academic research data accessible via the internet
- Retrieval of a specific information item is highly related to metadata including
  - established standards such as Dublin Core (ISO 15836:2009), IMDI (IMDI Part 1; IMDI Part 1 B), or OLAC (Simons and Bird 2008; Bird and Simons 2009),
  - newer specifications such as the Component Metadata Structure Initiative (CMDI, Broeder, Schonefeld, et al. 2011; Broeder, Windhouwer, et al. 2012) which allows both for documenting research information and querying it over the distributed repositories

ACCESSIBILITY OF PUBLICATIONS

- Another aspect of information infrastructure is the archival and accessibility of publications
- An in-house publication server can be a means for a research institution to retain both copyright and access control over information produced
- Open source implementations such as ePrints[1], or eSciDoc[2] often combine the functionality of publication servers with those of primary data repositories

[1] http://www.eprints.org/

[2] https://www.escidoc.org/

A QUESTION OF FORMATS

- A large portion of research data gets lost only shortly after the end of a project because of …
  - … hardware failures (lack of IT infrastructure) or insufficient metadata
  - … proprietary storage formats, for which the corresponding application is not available any more
- The decision for or against an open or proprietary data format is crucial when it comes to process and archive information

FORMATS FOR RESEARCH DATA

- Research data curated by an information infrastructure should be stored in open text-based formats
- Formats based on the open meta language XML such as TEI Guidelines or DocBook are quite common in academic research and can be defined by document grammar formalisms, allowing for on-the-fly-validation during the creation of instances
- 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 or CSS stylesheets
- This does not only prevent vendor lock-in, but significantly eases archival

TO SUM UP…

Information Infrastructure issues

To tackle the afore-mentioned Information Infrastructure issues regarding

- storage,
- accessibility,
- metadata, and
- formats

IT infrastructure and information infrastructure staff has to work together hand-in-hand

You may see a pattern here …

1 Introduction

2 IT infrastructure

3 Information Infrastructure

4 Legal issues

- Copyright
- Personal data protection

5 Conclusion

LEGAL ISSUES

Why do we have to talk about it?

Isn’t this talk about *digital* research infrastructure – i. e. technical things?

LEGAL ISSUES

Short answer

The best IT Infrastructure providing access to your Information Infrastructure is not enough if you are not allowed to share your research data

WHY LEGAL ISSUES ARE IMPORTANT

Research institutions are confronted with a number of legal questions, from which two key issues can be identified:

- Copyright issues, and
- personal data protection/privacy issues

COPYRIGHT 101

- Research data is often based on material contributed by external authors
- Primary data of text corpora for example often originates from non-academic sources
- In Germany, copyright law states that any work of literature, research or art (including software) is protected in its form (not the idea itself) – a sufficient level of creativity provided
- Changes to the work (including annotations added to a text) are only allowed by the copyright holder

COPYRIGHT AS A CONSUMER

- Although the German law does not contain the American concept of “fair use”, there are copyright restrictions (§§ 44a-63a UrhG) that apply to personal and scientific use, including citations and other use cases of copyrighted work (Mönch and Nödler 2006)
- Some of these restrictions apply only to “small groups of researchers” – a term that may not apply to open distributed research groups (Hoeren 2014, p. 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

COPYRIGHT AS A PRODUCER

- Research data for which a research institution holds the copyright (e. g., primary data that has been produced in-house) should be made available to others under a liberal license, e. g. an Open Access license such as the Creative Commons license[3]
- Creative Commons is a free license that consists of several license building blocks, such as
  - attribution to the original author (CC BY – minimal requirement),
  - NoDerivatives (CC ND),
  - NonCommercial (CC NC), and
  - ShareAlike (CC SA)

[3] http://creativecommons.org

COPYRIGHT AS A PRODUCER

- Apart from the easily-written CC license deeds, the license laundry symbols established in the CLARIN research group (Oksanen et al. 2010) provide a quick overview about license categories and respective licenses
- Recent extensions to these categories have been discussed by Kupietz and Lüngen (2014)

COPYRIGHT IN PUBLICATIONS

- Research institution's staff may agree to publish their works on the institution's publication server under an Open Access license (Degkwitz 2007) → in-house-policy
- While Open Access journals still sometimes lack some reputation compared to traditional journals (although both publication types employ peer review to ensure the journal's quality), they often have higher citation numbers (Stempfhuber 2009, p. 119)

PERSONAL DATA PROTECTION

- Privacy data issues may arise when living persons are involved in the creation process of research data, e. g. in voice (or video) recordings → a publication is only allowed if the rights have been granted (in written form) by the persons recorded
- For every collection of personal data, a register of processing operations (according to §4g, §§18 and 4e BDSG) has to be created, in which the type of personal information, its processing and the data protection measures are documented, amongst others
- In general, the data protection officer of the research institution should be involved as soon as possible

CONCLUSION

- Digital research infrastructures often combine aspects of IT and Information Infrastructure
- To address technical, financial, and legal aspects, different personell has to collaborate – at least in terms of technical issues, cooperations with other institutions can reduce costs and distribute work load
- Collaboration should start *before* a projects begins

INSTITUT FÜR
DEUTSCHE SPRACHE

THANK YOU FOR YOUR ATTENTION!

stuehrenberg@ids-mannheim.de

Member of the
Leibniz Association