Metadata & Repositories

Matej Őurčo @ ACDH Tool Gallery 2.1, 2016-03-16
the path

describe – deposit – share – publish  |  discover – access – cite
interdependencies

purpose

data model – format – system

usability availability
digital preservation

formal endeavour to ensure that digital information of continuing value remains accessible and usable. (Wikipedia)

make resources/datasets
  ▶ persistent
  ▶ discoverable
  ▶ understandable
  ▶ accessible
  ▶ citable
“data about data”
administrative/structural/technical/provenance/descriptive
Location:
- separate: XML-files (CMDI, DC), Databases
- embedded: JPG, TEI, ...
Metadata/data/annotation distinction?
Especially with RDF or in relational databases
Interoperability
Be able to exchange data across systems (keeping the semantics)
Single sourced
Use the most comprehensive format and derive the others
explicate the model
DDL, DDT, ODD, XSD
metadata formats

- **DC** – dublincore (elements/ DCMI terms)
- **METS/MODS** (LoC)
- **ALTO** - Analyzed Layout and Text Object (technical metadata for OCR, LoC)
- **CMDI** – Component Metadata Infrastructure (CLARIN)
- **EDM** - Europeana Data Model
- **DCAT** – Data Catalog Vocabulary (w3c)
- **ORE** – Object Reuse and Exchange
- **EAC-CPF, EAD, EAG, ISAD, ISAR, ISDIAH** - Archival Holdings
- **DDI** – Data Documentation Initiative (DDI) statistical and social science data.
- ...

Vocabularies / Classification schemes

- **SKOS** – Simple Knowledge Organisation System (w3c) as lingua franca
metadata authoring

- relational database
  - oXygen
- generic XML editors
  - oXygen
- specialized tools
  - ARBIL, COMEDI
- repository submission
  - PHAIDRA
  - LINDAT
metadata workflow - harvesting, curation, publishing
save the data!

- **store/persist reliably, long-term**
  - bit-stream preservation -> redundancy (LOCKSS), fixities
  - ensure renderability
  - counter format and media obsolescence through migration

- **allow**
  - Structured datasets (collections, relations)
  - Custom metadata (but not “anything goes”)
  - Flexible data formats (but not “anything goes”)

- **how long is long-term?**
  - what needs to be stored long-term
  - courage for cassation

from: http://archaeologydataservice.ac.uk/advice/DepositingData#section-DepositingData-HowToDeposit
Repositories

- Requirements
  - OAIS reference model (ISO 14721:2012)
    defined workflow, roles, structure (SIP, AIP, DIP)
  - DSA - Data Seal of Approval (16 guidelines)
  - CLARIN B Centre Assessment

- Roles
  - Data Producer
  - Collection Manager
  - Data Consumer
repositories

- **Software:**
  - Fedora, DSpace, CKAN, ...

- **Services:**
  (institutional / domain-specific / infrastructural)
  - [GAMS](#)
  - [PHAIDRA](#)
  - [epub.oeaw](#)
  - [ads](#) – archaeology data service
  - [CLARIN Depositing Services](#)
  - [Datahub](#) by Open Knowledge Foundation – “give your data a home” (10.695 datasets)
  - [Figshare](#) – “credit for all your research”
  - [Re3data](#) – Registry of research data repositories
  - [COAR](#) – Confederation of Open Access Repositories
not an (digital) object?

- lots of data in relational databases
- regular backup is necessary but not sufficient for long-term
  - persistent identifier and descriptive information is missing
- persistency options:
  - SQL dump + description into repository
  - generic serialisation
    - [SIARD](#) – Software Independent Archiving of Relational Databases
    - [D2RQ](#) – convert relational databases to RDF
  - high-level application export (XML, RDF)
- RDF/LOD
  - easy to serialise
  - self-descriptive, self-contained
  - store dump in repository
  - [datahub](#), [linghub](#)
  - (often used in repositories for metadata/relations of the objects)
• disseminate metadata over many channels linking back to data in the repository

• Metadata catalogues / Aggregators
  ▶ CLARIN VLO (~ 1 Mio. Language resources)
  ▶ Europeana (52 Mio. Objects ?)
  ▶ recherche-isidore.fr
  ▶ OpenAIRE (13 Mio. pubs, 17.000 datasets, ~ 6.500 repos)
  ▶ OLAC - Open Language Archives Community
  ▶ JSTOR - http://www.jstor.org/
  ▶ narcis.nl@DANS – (1,23 Mio. publications, ~150.000 datasets, 1710 enhanced publications)

• OAI-PMH – protocol for metadata harvesting
  ▶ provider exposes metadata via endpoint
  ▶ harvester regularly fetches metadata
  ▶ one Registry of data providers
raw data vs. search endpoint vs. application vs. visualisation
- direct (persistent) link vs. CD-ROM per mail
- landing page
- Federated Identity
- Restrictions (License and availability)
  - CLARIN License categories (VLO)
  - Open Research Data Pilot
    - H2020 projects required to make the produced data available

CAN I USE IT?
- http://www.europeana.eu/portal/search
- Only With Permission (24,924,042)
- Yes With Attribution (18,969,044)
- Yes With Restrictions (8,159,444)

http://beta-vlo.clarin.eu

The following information regarding availability and/or licencing has been extracted:
- Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
- PUB language resources can be distributed publicly. The distribution data protection issues.
- Attribution, i.e. acknowledgement of authorship, is required
- Content is available only for non-commercial purposes

Show all available licence/availability information

Please verify this information at the original source or contact the original producer.
- may be temporary
- for selected people
- aka „cloud services“
- commercial
  - Dropbox
  - Google-drive
- institutional
  - Oeawcloud (based on owncloud)
- research infrastructures
  - DARIAH-DE Repository
Persistent Identifiers (PID)
- Handle.net, DOI, ARK
  http://hdl.handle.net/11858/00-1734-0000-0009-FEA1-D

Activities
- DataCite – DOI(PID) for datasets
- Thor - integration between articles, data, and researchers across the research lifecycle
- RDA Working Group on Data Citation Dynamic Data Citation

Cite-helper LINDAT: Czech WordNet 1.9 PDT

Please use the following text to cite this item or export to a predefined format:

Pala, Karel; Čapek, Tomáš; Zajičková, Barbora; Bartůšková, Dita; Kulková, Kateřina; Hoffmannová, Petra; Bejček, Eduard; Straňák, Pavel and Hajíč, Jan, 2011, Czech WordNet 1.9 PDT, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, http://hdl.handle.net/11858/00-097C-0000-0001-4880-3.
data centre - Datenzentren

- **DHd WG** (convenor: Patrick Sahle)
- not just an archive, not just a computing centre
- comprehensive support over whole project duration
  “archiving begins with the project conception!” (Johannes Stigler)
- harmonized set of supporting services
- domain-specific expertise + technical know-how
- **advice, guidance, consulting**
  - ads/advice
  - IANUS@DAI - nationales Forschungsdatenzentrum, Empfehlungen
  - Anforderungen an Repositorys für Dokumente

From: http://archaeologydataservice.ac.uk/advice/DepositingData#section-DepositingData-HowToDeposit
Thank you!

Questions?

Matej Ďurčo @ ACDH Tool Gallery 2.1, 2016-03-16
Text
a dynamic, evolving, continuous process
a graph-like ever-growing body of data

WWW
Sources
Annotation / processing
Visualization

Automatic
Hybrid
Manual
Cumulative
Versioning

crowdsourcing
linguistic
semantic

Digital Edition
a research question driven
dynamic configurable VRE
Component Metadata Infrastructure: Profiles/Components/Elements/Concepts
DataSheet editor

http://geobrowser.de.dariah.eu/edit/

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Description</th>
<th>Longitude</th>
<th>Latitude</th>
<th>TimeStamp</th>
<th>TimeSpan:begin</th>
<th>TimeSpan:end</th>
<th>GettyID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wien</td>
<td></td>
<td></td>
<td>16.3687</td>
<td>48.2167</td>
<td></td>
<td></td>
<td></td>
<td>7003321</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
save the data!

Repositories
Requirements on online availability

Varying combinations of:

▸ full-text search
▸ semantic search (search for persons, places, concepts, search by classifications)
▸ full-view (e.g. text and facsimile of individual pages)
▸ specialized visualizations (temporal, spatial, graph, statistical data)
▸ raw data available for download
▸ stable references to resources and resource fragments

BUT before publication: collaborative editing → VRE!
Repositories

- CLARIN Centre Vienna / Language Resources Portal (FEDORA-based)
  - Instance of GAMS + client: Cirilo by Uni Graz
- vs. epub.oeaw
  - run by Academy Press
  - Software Hyperwave
  - Mainly for publications, but also some structured data (lexicographic databases/apps)
  - mirrored to Austrian National Library and PORTICO

BUT.

- Relational DBs
  - adlib – commercial software for Archives, Libraries and Museums (Axiell company)
  - Custom Django applications (APIS, DEFC), tokenEditor
- RDF-data -> Triple store
- Corpora -> SketchEngine, Solr, ddc
- All on ARZ Servers – NetApp - regular snapshots, 2x replication)
Proposed reshaping of the workflow with central administrative dashboard
Current view of the overall architecture

Hybrid/integrated search

Knowledge aggregation

Index 1

Index 2

...

Directory service of providers/people

Keep identities of the who, what in the infrastructure

Minimal identity metadata, part-of, agency about:

Datasets, metadata sets, software, services, mappings, and more

Joint resource registry

Mappings, Resources integration

Aggregation mechanism

Content cloud

Common design requirements

Resources: institutional research environments