European Language
Resource Coordination
Connecting Europe Facility

Best practice for the future: Capitalize on your valuable data

Best Practice in Data Management

David Lewis, Associate Director ADAPT Centre, Trinity College Dublin

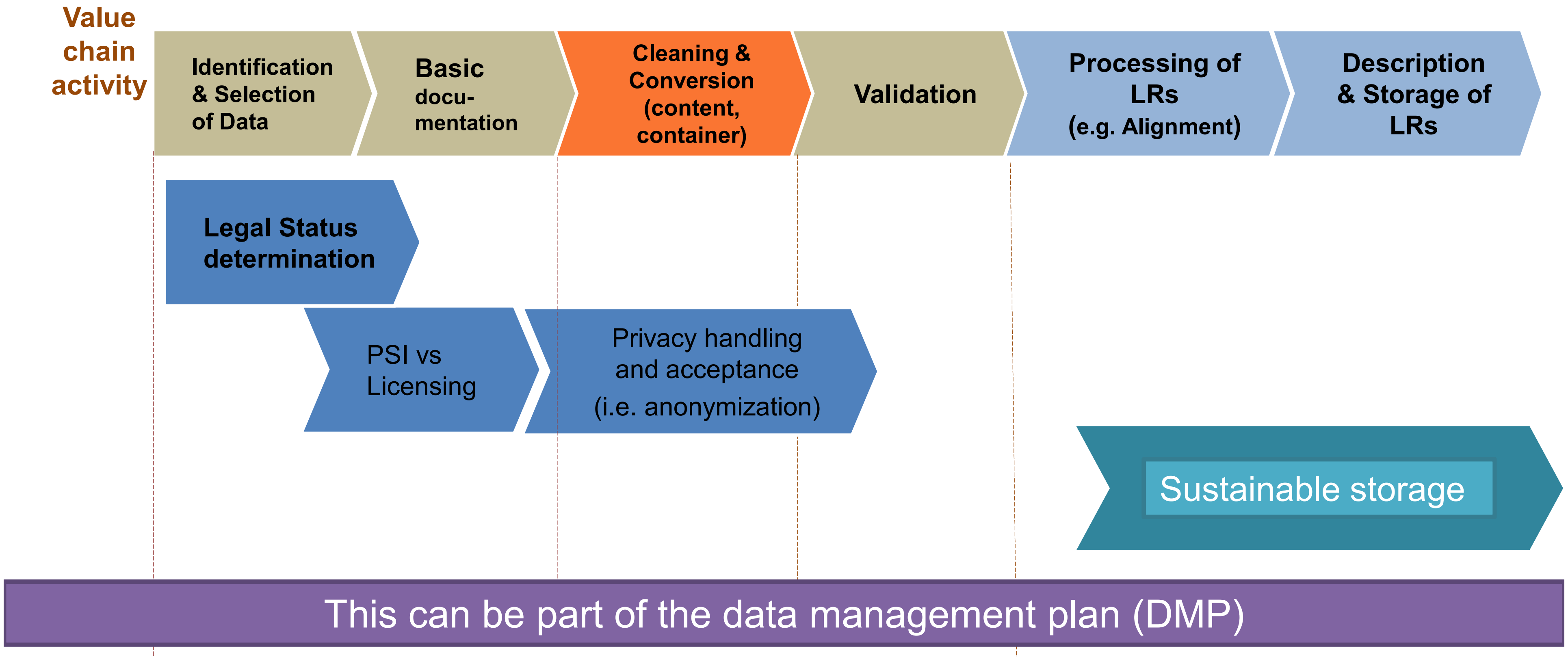
My data in the future

- Now that I know the value of data, what should my plans be?
- What are the best ways to collect, maintain, archive and re-use my data
- In particular how can I use it for improving MT performances?

Five best practices for data management

1) Analyse all phases of data development
2) Based on 1), create a data management plan
3) Take all relevant aspects for the DMP into account
   - Legal, tasks of data, formats, publication as PSI, …
4) Consider data sustainability
   - Data specification, production, validation, sharing & distribution, maintenance & preservation
5) Use the Web as an additional publication channel
   - Machine readable standardised data on the Web; potentially usage of linked data standards

Main phases of data development


Value chain activity
Identification & Selection of Data
Basic docu-mentation
Cleaning & Conversion (content, container)
Validation
Processing of LRs (e.g. Alignment)
Description & Storage of LRs
Legal Status determination
PSI vs Licensing
Privacy handling and acceptance (i.e. anonymization)
Sustainable storage

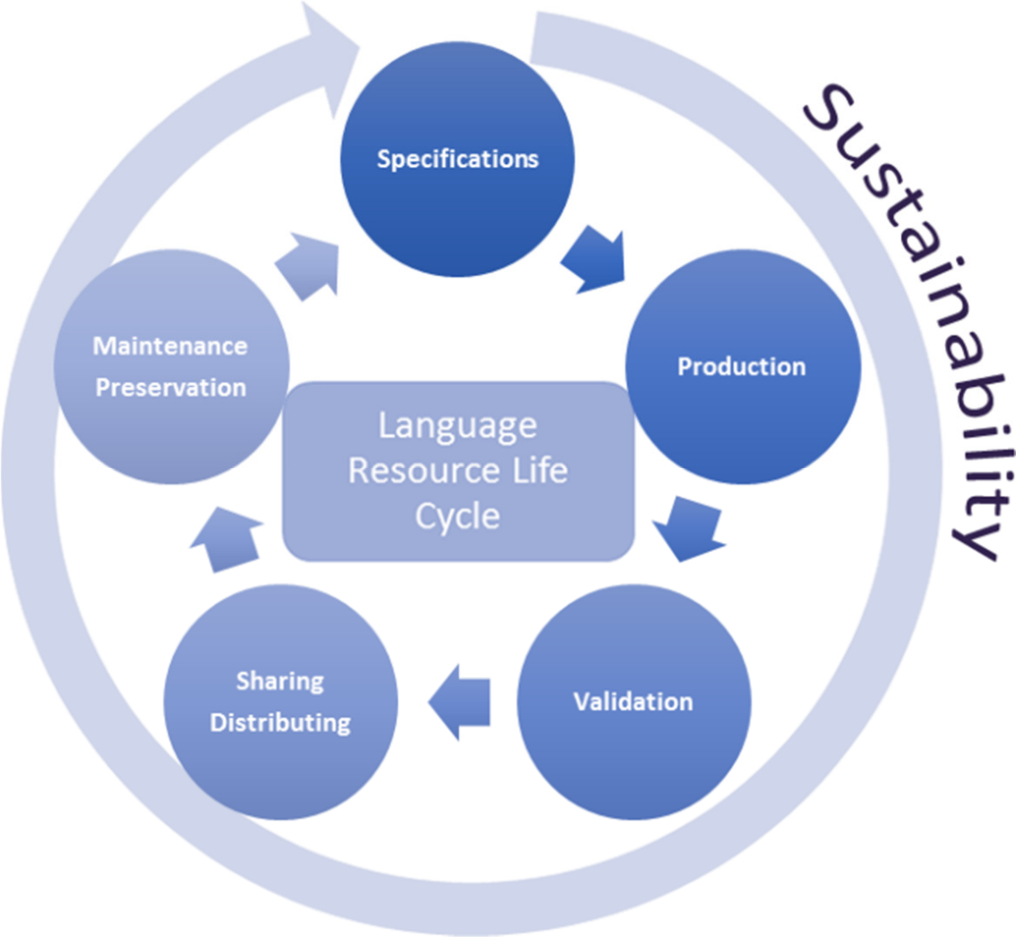
This can be part of the data management plan (DMP)

Concerns in creating a DMP

- Anticipate all potential legal issues
  - Ensure that your data IPRs are cleared
  - Ensure that the producing parties adhere to your right “ownership” (e.g. relations with LSP: ensure you keep all rights)
  - Ensure that all produced intermediary documents are yours (e.g. translation memories)
  - Check the privacy issues in advance and plan for anonymization if necessary
- Define your management plan with respect to the task
  - This has to account for the main goal (e.g. document writing, doc translation, etc.)
- Plan for repurposing (from documentation to LRs)
  - Request data in a usable format (not only PDFs but also TMX/Word/XML/TXT)
  - Make sure that your data uses up-to-date medium (no CDs?)
- Foresee for future publication and sharing as Public Sector Information (PSI)

European Language Resource Coordination
Connecting Europe Facility

Key elements of a Data Management Plan



Key elements of a Data Management Plan

- Specifications
  - Ensure that the original documents are described
  - Ensure that your needs are described
  - Anticipate what you can get as valuable resources (a side effect)
- Production
  - Whether internal or outsourced, check that the tools used are compatible with your needs and beyond (e.g. CAT, MT, etc.)
  - Ask for the list of tools and production software
  - Check if you can get texts in the multiple languages aligned to each other
  - Keep a clear documentation of the data being produced (meta-data)

Key elements of a Data Management Plan

- Validation
  - In addition to your quality control, you may want to use some of the validation tools (lexical coherence, syntactic analysis, etc.)
- Sharing/distribution
  - Ensure your data falls within the PSI directive as transposed in your country
  - If not, foresee an open and permissive licence
  - If privacy is an issue, plan necessary procedures to handle these
- Maintenance/preservation
  - The best option is often to partner with a data centre
  - See how ELRC can assist you
  - There is also the option of national open data portal
  - Only “putting” data on the web is not a sufficient option (referencing?) – but an additional one (see following slides)

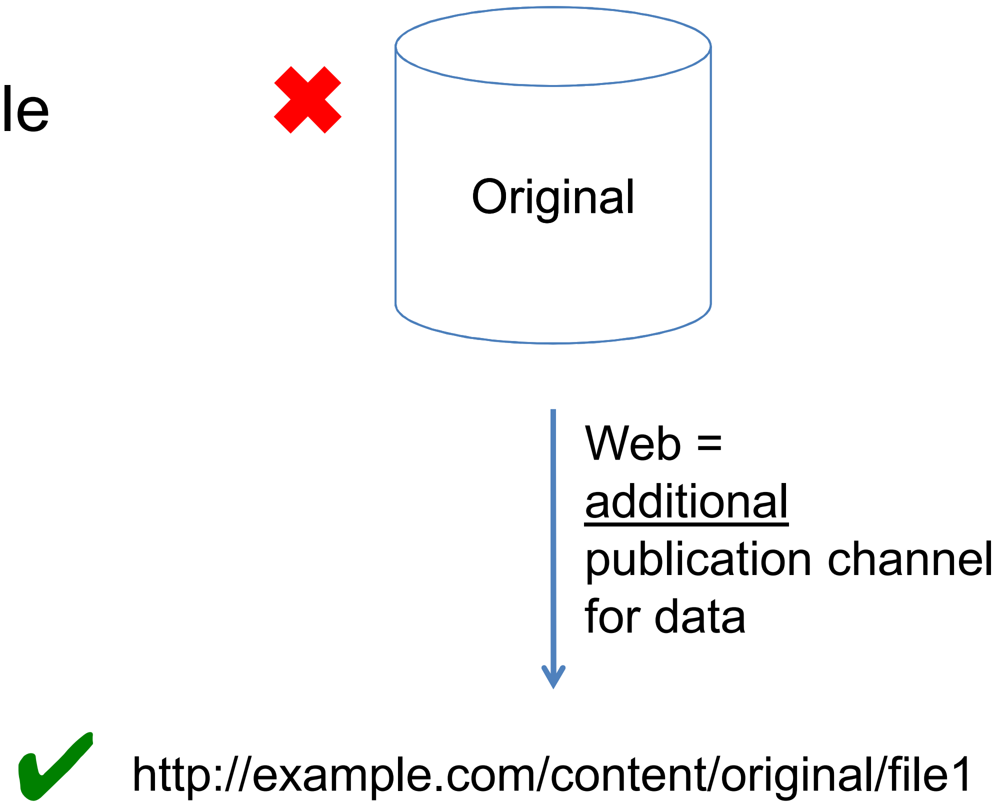
The Web as an additional publication channel: towards 5 star linked data*

1. Put data on the web
2. Provide machine-readable data
3. Use non-proprietary formats
4. Use RDF standards
5. Provide linked data

* See http://www.w3.org/DesignIssues/LinkedData.html

The Web as an additional publication channel: towards 5 star linked data

1. **Put data on the web**
2. Provide machine-readable data
3. Use non-proprietary formats
4. Use RDF standards
5. Provide linked data



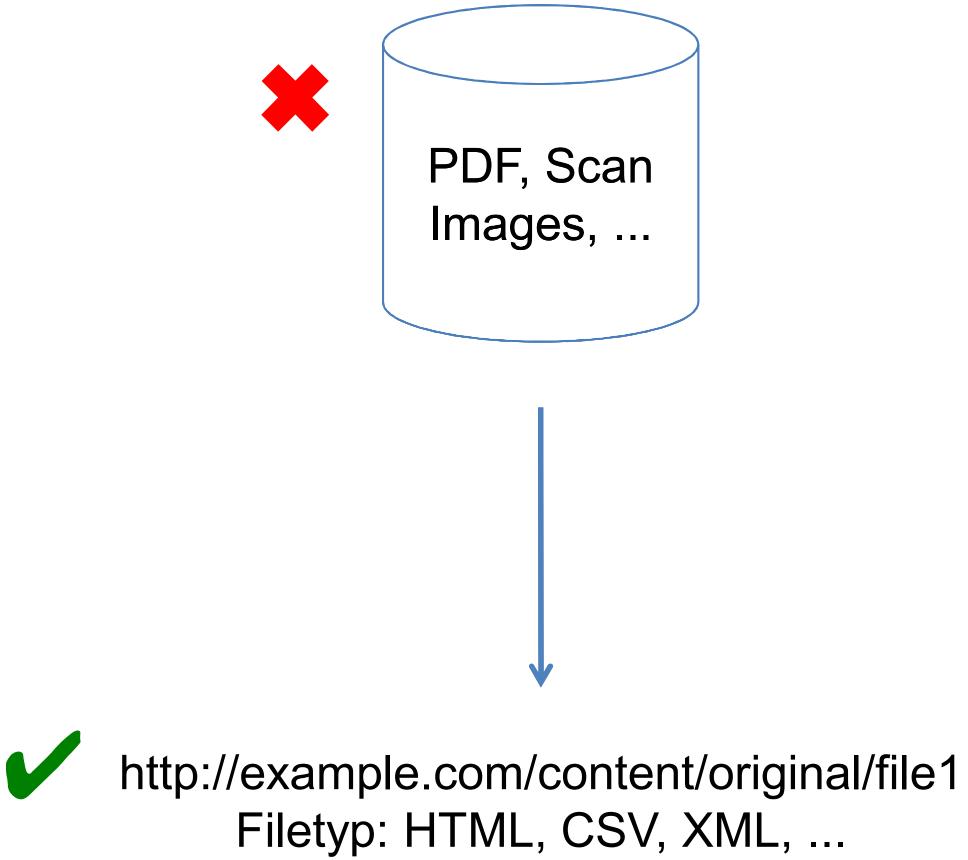
✗ Original

Web = <u>additional</u> publication channel for data

✓ http://example.com/content/original/file1

The Web as an additional publication channel: towards 5 star linked data

1. Put data on the web
2. **Provide machine-readable data**
3. Use non-proprietary formats
4. Use RDF standards
5. Provide linked data

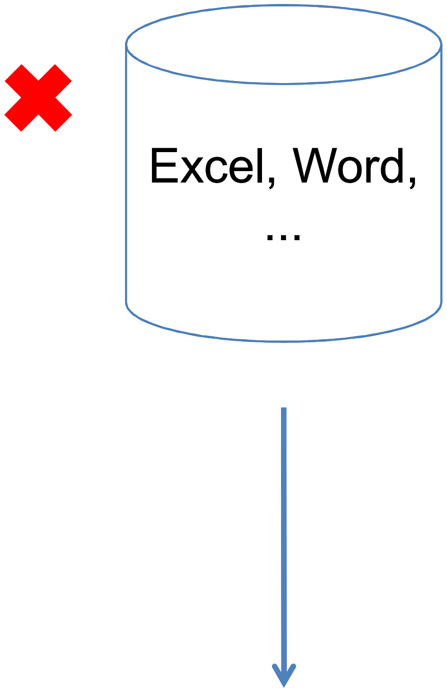


✗ PDF, Scan Images, ...

✓ http://example.com/content/original/file1
Filetyp: HTML, CSV, XML, ...

The Web as an additional publication channel: towards 5 star linked data

1. Put data on the web
2. Provide machine-readable data
3. **Use non-proprietary formats***
4. Use RDF standards
5. Provide linked data



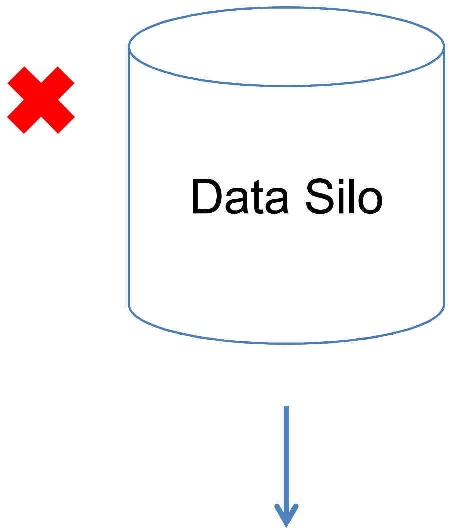
✖ Excel, Word, …

✔ http://example.com/content/original/file1
Filetyp: TBX, TMX, XLIFF (nicht verlinkt) >
RDF (verlinkt)

*Standards = minimize costs through
- ✓ Improved interoperability
- ✓ Use of Open-Source Tools
- ✓ Availability of „royalty-free“ technologies
- ✓ Open formats enable extensions

The Web as an additional publication channel: towards 5 star linked data

1. Put data on the web
2. Provide machine-readable data
3. Use non-proprietary formats
4. **Use RDF standards***
5. **Provide linked data**



Data Silo

✔

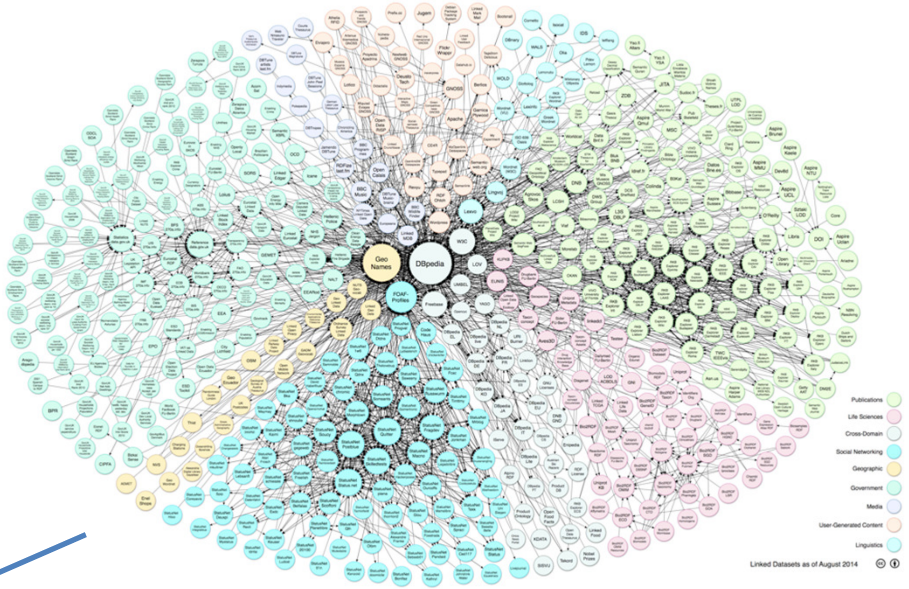
- Linked Data Cloud and
- Linguistic Linked Data Cloud
- No replacement for TBX, TMX, XLIFF etc.; just used as an additional publication channel

* RDF = "Resource Description Framework", basis for Linked Data Technologies

European Language Resource Coordination
Connecting Europe Facility

The Web as an additional publication channel: towards 5 star linked data

1. Put data on the web
2. Provide machine-readable data
3. Use non-proprietary formats
4. **Use RDF standards**
5. **Provide linked data**



Registration and Index services:
https://datahub.io/
http://linghub.lider-project.eu/

- Linked data cloud
- Decentralised, like the web itself
- Success stories in several domains, e.g. medicine, libraries, archives
- Relevant also for linguistic data – the linguistic linked data cloud