Sustainable operability: Keeping complex linguistic resources alive


Menzo Windhouwer


Alexis Dimitriadis

Sustainability of text resources

- All electronic resources are accessed with the mediation of appropriate **software.**
- Text editors, and web browsers, are **generic:** they can be used, with satisfactory results, with any document in a supported format.
- Documents are (more or less) **portable** across software that supports their storage format.

- Sustainability can be safeguarded by relying on documented, standard formats for their encoding.

The problem with databases

- Databases are **not portable** in the sense that text documents are:
- The data and relational structure of databases can be stored in (semi-)standard SQL format, or exported to other formats.
- But databases are typically accessed through a custom-made user interface. Preserving the data, therefore, does not preserve the **complete resource.**
- In this talk, we focus on **(typological) databases.**

Operability of complex resources

- The general problem: **Complex resources depend on custom software.** Without the software, the resource is not usable and hence not truly preserved.
- We will call a resource **operable** if suitable access or management software (operating software) exists for it.
- While all electronic resources depend on software for their operability, **complex resources** are particularly vulnerable because they **lack an economy of scale**.

Outline

- The problem of sustainable operability
- Sustainable operability of typological databases
- The IDDF architecture
- The Typological Database System
- The TDS Curator project

Next

- The problem of sustainable operability
- Sustainable operability of typological databases
- The Typological Database System
- The IDDF architecture
- The TDS Curator project

Typological databases

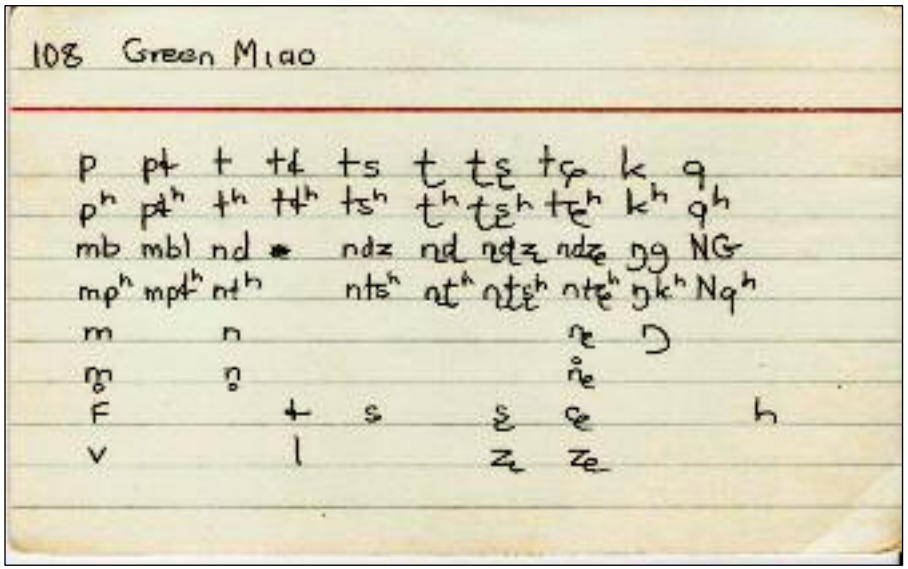
- Contain high-level, summary information about selected phenomena in a large number of languages.
- May include example sentences with interlinear gloss annotations.
- Are implemented on a variety of software platforms (Filemaker, MS Access, MySQL, 4th Dimension, Excel spreadsheets, custom software), and may or may not have a web interface.

Databases are diverse:

- Original schema snippet from TDN database

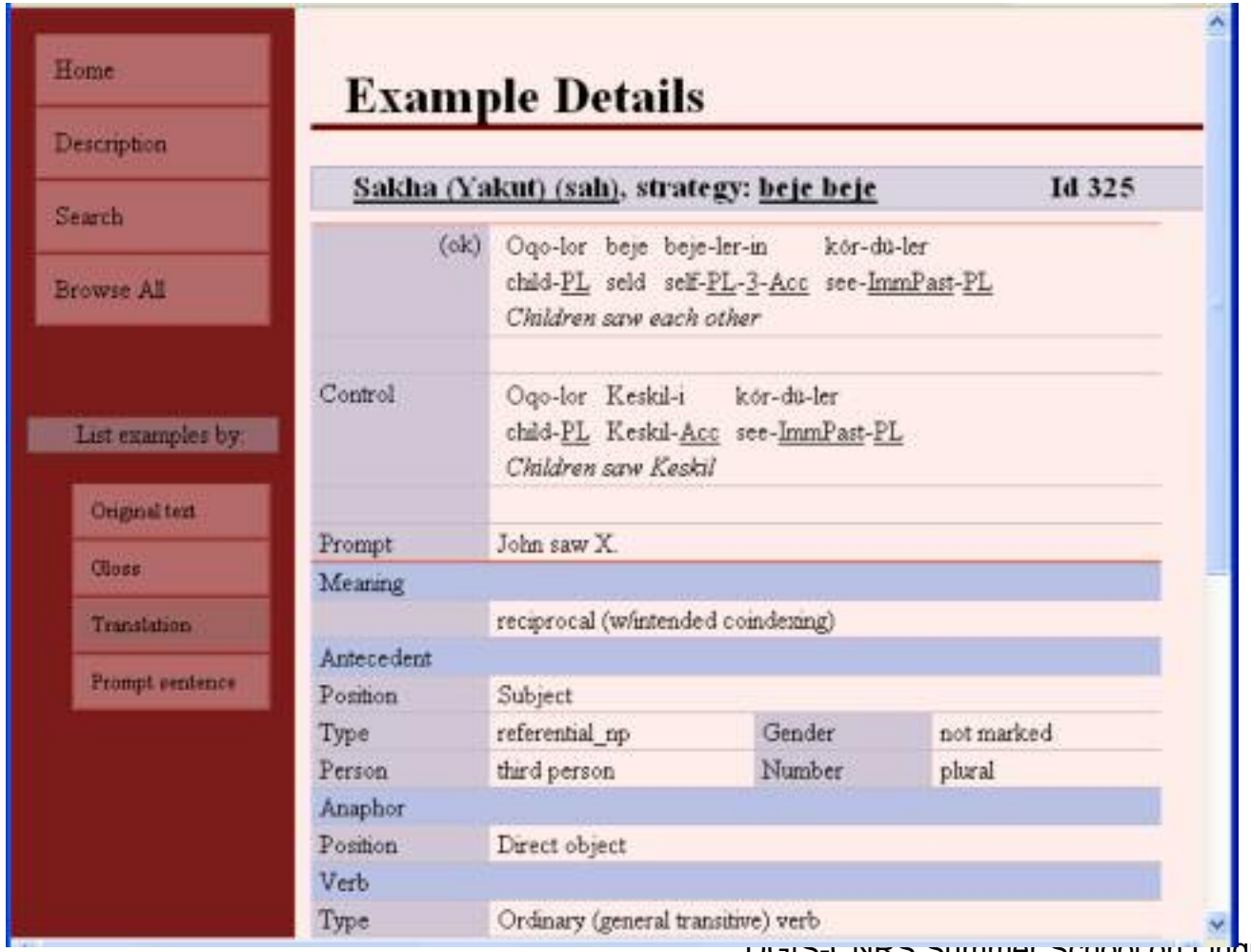
Field	Values	Metadata
V105	0, 1, 9, 99	ATTRIBUTIVE ADJECTIVES ARE RELATIVE CLAUSES
V168	0, 1, 9, 99	PRED LOC = ZERO + LOC PP
V204	0, 1, 9, 99	PRED ADJ = COP VS. PRED LOC = VERB (NONCOP)

- Phoneme inventory (SPIN database)



Databases are diverse:

- User interface snippet for the Anatyp database



Typological databases – their fate

Completed databases are subject to the usual perils:

- Gradual obsolescence of db software, OS, or hardware.
- Sudden disappearance due to incompatible software updates, retirement of legacy servers, or hardware failure.
- Gradual fall into unusability, with the dissipation of the insider knowledge needed to utilize a poorly documented database.

A data dump is insufficient

- Why not just export a database's tables in some standard format (tab-separated Unicode text, or even a dump in "standard" SQL)?

- This would still be deficient in
  1. Completeness of content and documentation
  2. Operability

Completeness

- The meaning of table contents, and their interrelationships, are not explicitly given in the data tables
- D… nal s…
  - …nly in
- In the OAIS terminology, data tables alone are rarely "independently understandable".

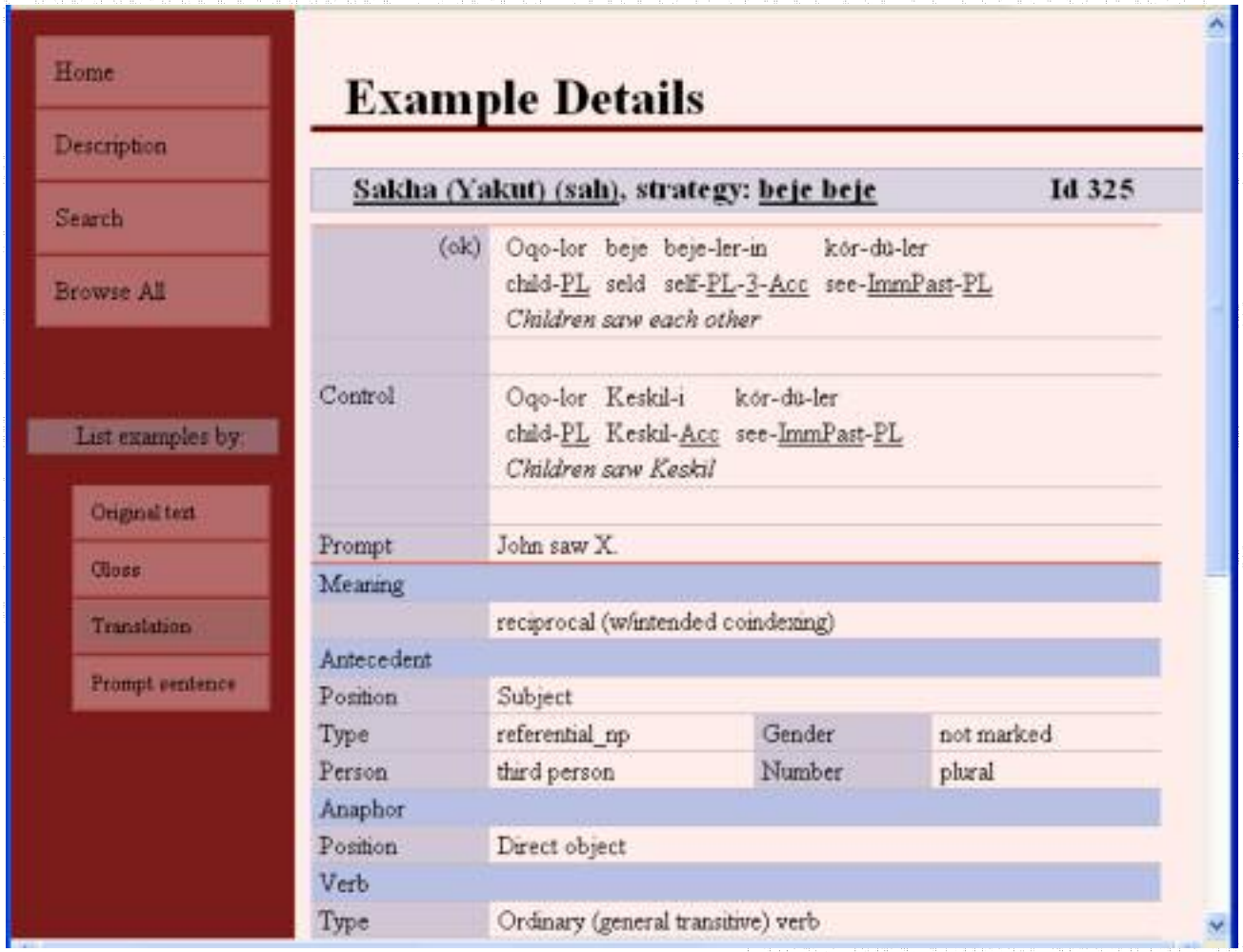
Completeness example: TDN database

Field	Values	Metadata
V105	0, 1, 9, 99	ATTRIBUTIVE ADJECTIVES ARE RELATIVE CLAUSES
V168	0, 1, 9, 99	PRED LOC = ZERO + LOC PP
V204	0, 1, 9, 99	PRED ADJ = COP VS. PRED LOC = VERB (NONCOP)

Operability

- Th
  su
- Th
  en
  - se
  - y select,
  - which

Presentation example: Anatyp database



Our approach to sustainable operability

1. Map resources to a **sufficiently rich** format at time of archiving.
2. Maintain **generic software** that can provide browsing and query access to all archived resources in an application domain.

Next

- The problem of sustainable operability
- Sustainable operability of typological databases
- The IDDF architecture
- The Typological Database System
- The TDS Curator project

The Integrated Data and Documentation Format

- Data, structuring information and documentation are combined into an integrated, XML-based standardized format, the Integrated Data and Document Format (IDDF).
- Software is provided that can manage IDDF-encoded resources in a generic way, just as a text editor or corpus tool can manage arbitrary conforming resources.
- New generations of management software can be provided in the future, utilizing the self-describing nature of the IDDF and an economy of scale.

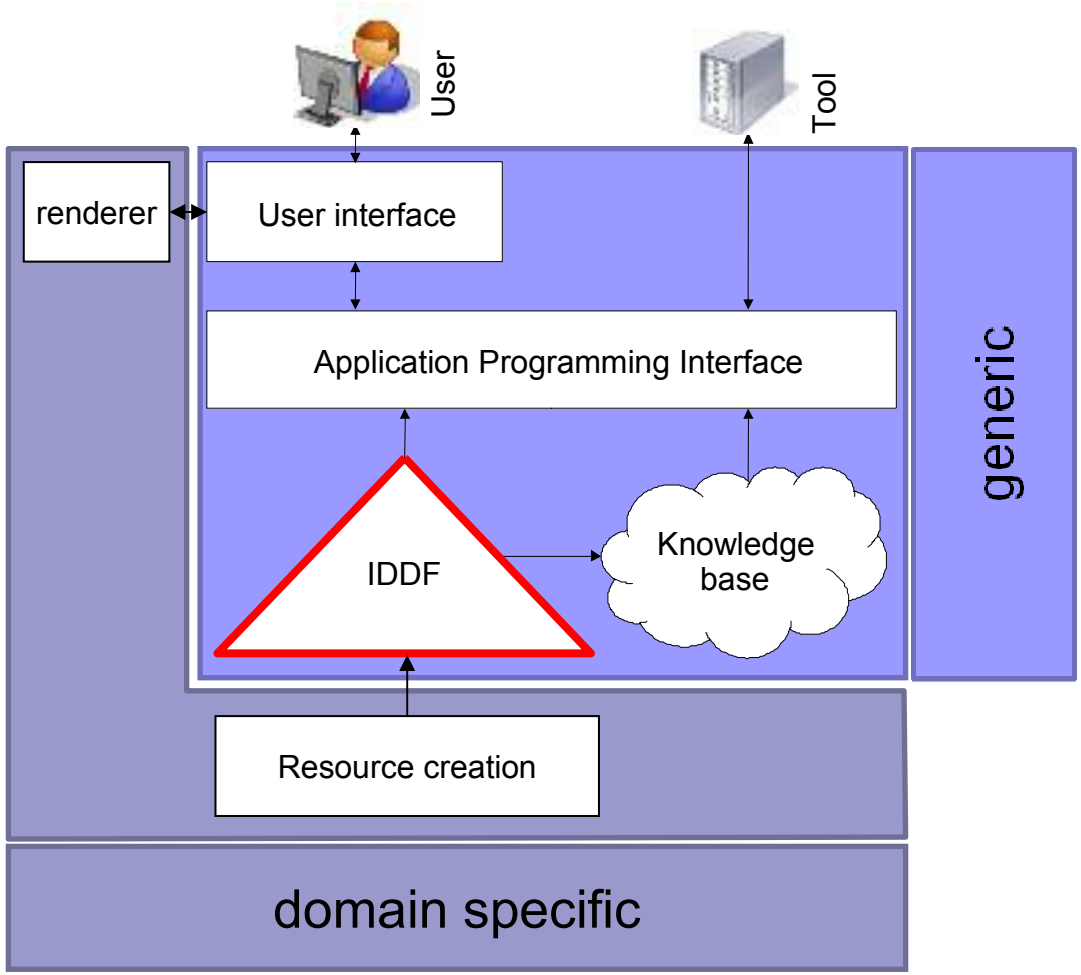
IDDF structure

- Two major sections:
  1. Metadata section:
     - provides the (loose) data schema
     - documents the elements in the schema
  2. Data section:
     - contains the actual data
- Hierarchical, semi-structured data model
- Network of hierarchical units, a.k.a. semantic contexts

IDDF: metadata

- For each data element:
  - A label and a description
  - One or more links
    - to other elements
    - to external resources, e.g., a knowledge base
  - Data types:
    - A semantic data type for the element, e.g. UPPC
    - A semantic (key) value data type, e.g. interlinear glossed text tier
  - An (partial) enumeration of possible values:
    - The literal (key) value
    - A label and a description
    - One or more links
      - to other elements
      - to external resources

IDDF: system architecture



Next

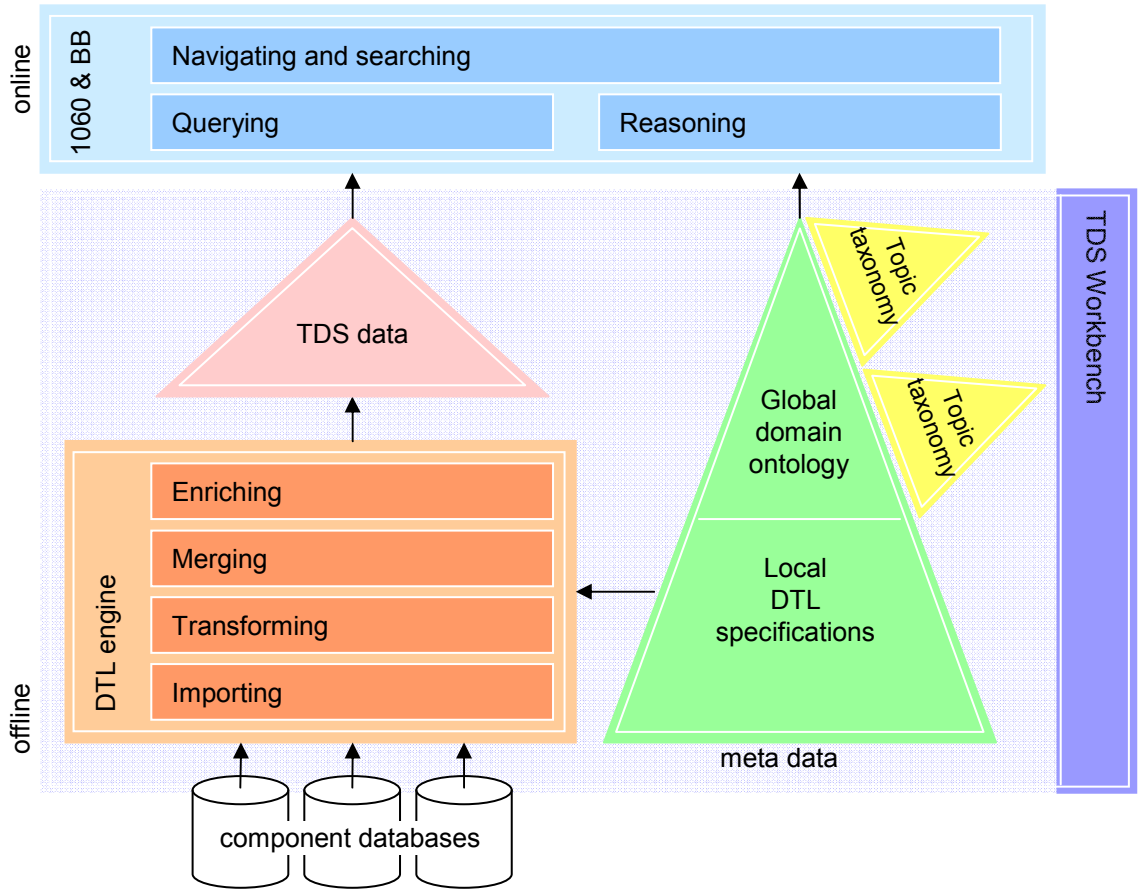
- The problem of sustainability operability
- Sustainable operability of typological databases
- The IDDF architecture
- The Typological Database System
- The TDS Curator project

The Typological Database System

- The Typological Database System (TDS) provides integrated access to multiple, independently created typological **databases.**
  - Provide an interface that will help users **find** relevant data.
  - Allow users to **interpret** the data they are presented with.
- The system behaves, as much as possible, as a single database.

  Various differences between the component databases must be dealt with.

TDS: system architecture



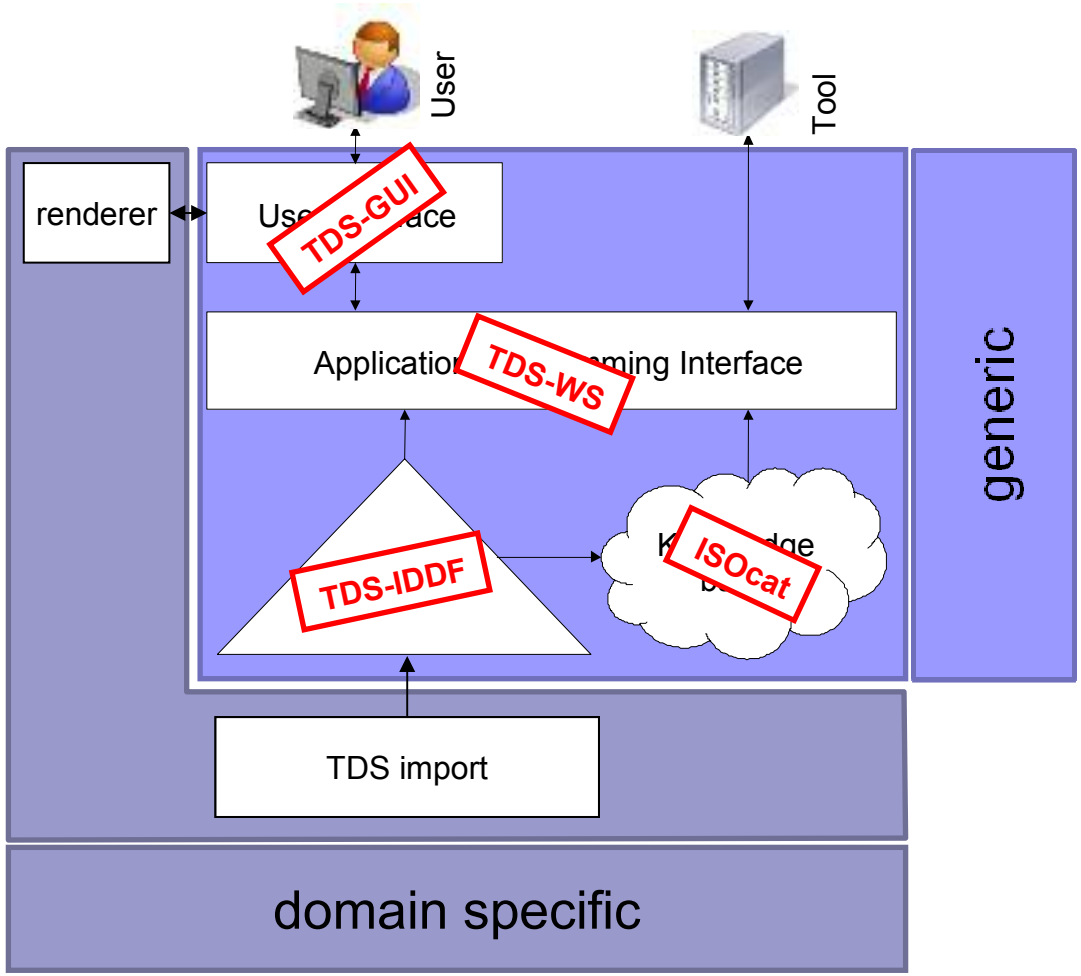
Next

- The problem of sustainable operability
- Sustainable operability of typological databases
- The IDDF architecture
- The Typological Database System
- The TDS Curator project

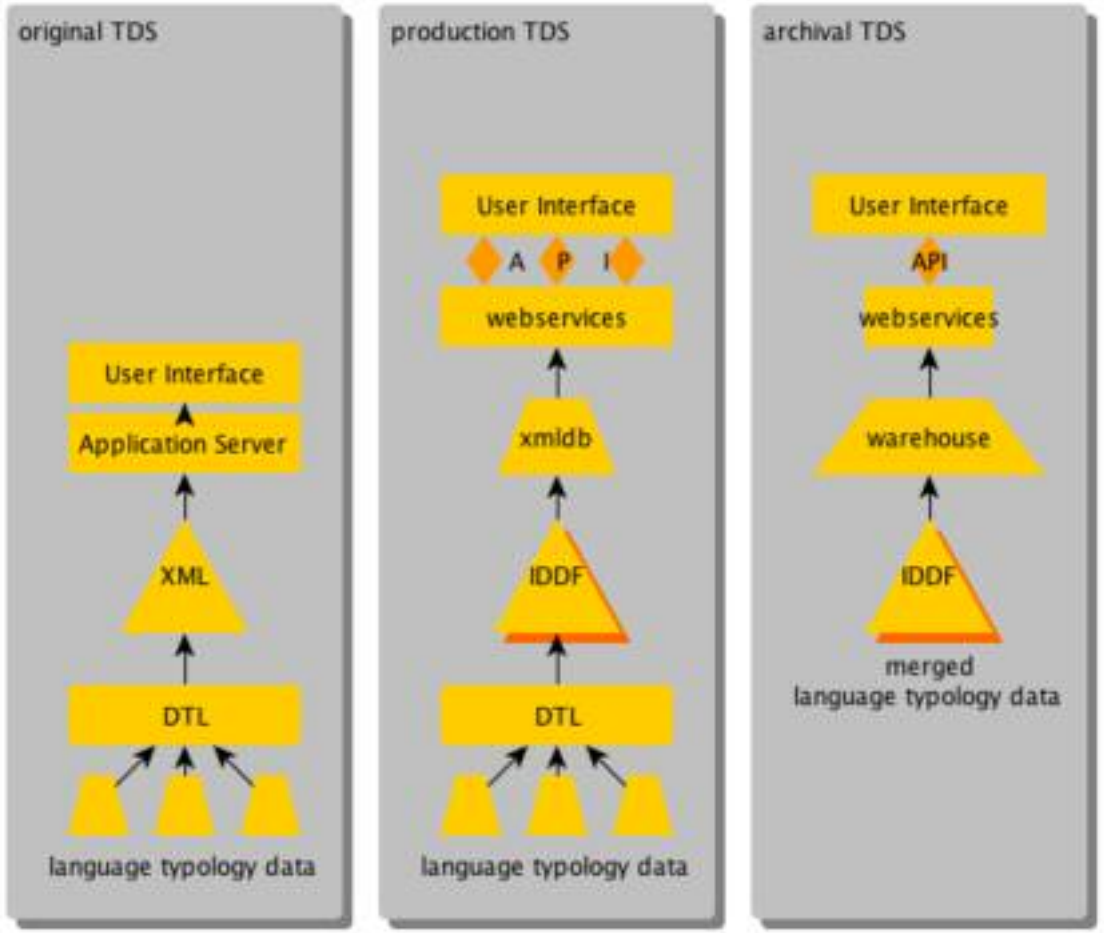
The TDS-Curator project

- CLARIN-NL Call 1 project
- “TDS Curator will make the TDS into a sustainable service that conforms to CLARIN infrastructural requirements.”
- Partners:
  - Utrecht University
  - DANS
  - Max-Planck-Institute for Psycholinguistics
- May – November, 2010

Work packages



IDDF-based sustainable operability



Summary

- Sustainable operability is a challenge for complex resources like typological databases
- A corner stone of a solution is a generic format that is rich enough to allow operability by generic tools
- The Typological Database System provides a promising architecture, which has been applied to more then a dozen of typological databases
- The propriety data model of the TDS can be turned into an open format, i.e. the Integrated Data and Documentation Format
- In the CLARIN-NL TDS Curator project this more generic setup will be realized

- It will be interesting to also use this generic system outside the domain of linguistic typology or even linguistics

http://languagelink.let.uu.nl/tds/

Thanks for your attention!

IDDF: top-level structure

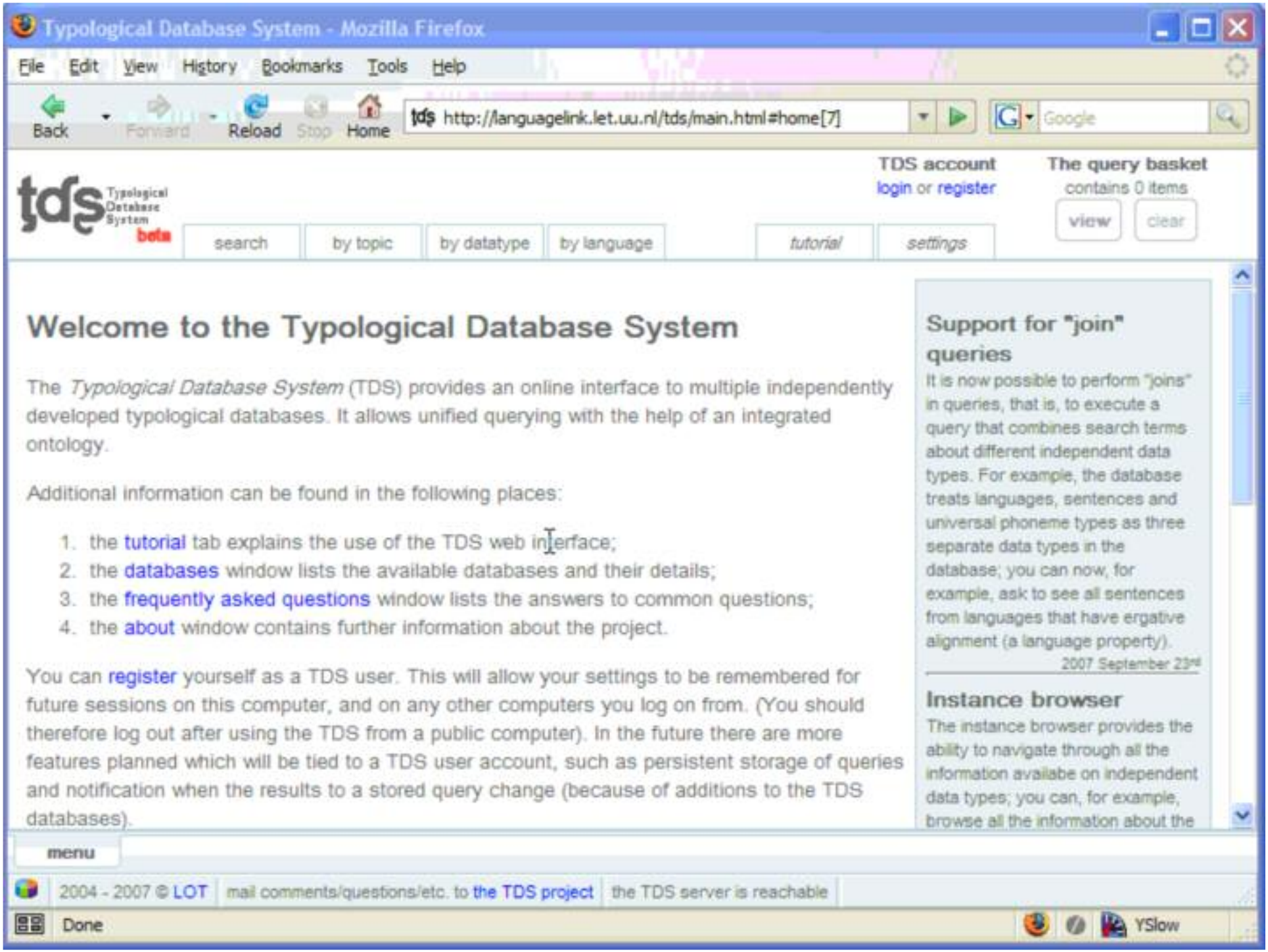
```
<iddf:warehouse xmlns:iddf="http://.../ns/iddf">
   <iddf:meta>
       <iddf:scope id="tds" type="warehouse">
              …
       </iddf:scope>
       <iddf:notion id="n1" name="language" scope="tds"
                              type="root" key-datatype="enum">
              <iddf:label>Language</iddf:label>
              <iddf:description>
                     One of the world's languages
              </iddf:description>
              …
       </iddf:notion>
       …
   </iddf:meta>
   <iddf:data xmlns:tds="..." ...>
       <tds:language iddf:notion="n1" key="...">
              …
       </tds:language>
       …
   </iddf:data>
</iddf:warehouse>
```

IDDF: metadata example

```
<iddf:notion id="n7" name="vowel" scope="SyllTyp">
   <iddf:label>Vowel</iddf:label>
   <iddf:description>
        Is the segment a vowel?
   </iddf:description>
   <iddf:link type="concept" rel="as“ href="...owl#vowel"/>
   <iddf:link type="concept" rel="to“
                          href="...owl#vocalicFeatureNode"/>
   <iddf:values datatype="enum">
        <iddf:value>
              <iddf:literal>+</iddf:literal>
              <iddf:description>
                    The segment is a vowel.
              </iddf:description>
        </iddf:value>
        …
   </iddf:values>
</iddf:notion>
```

IDDF: data example

```
<iddf:data xmlns:tds="…/ns/iddf/tds" … >
   <tds:language key="l-iso-tba"
                iddf:notion="n1" iddf:sources="SyllTyp UPSID">
        <tds:identification
                iddf:notion="n2" iddf:sources="SyllTyp UPSID">
                <tds:name
                iddf:notion="n3" iddf:sources="SyllTyp UPSID">
                        <iddf:value srcs="SyllTyp">
                                Wari’ (Tubar&#227;o)
                        </iddf:value>
                        <iddf:value srcs="UPSID">
                                Huari
                        </iddf:value>
                </tds:name>
                …
        </tds:identification>
        …
   </tds:language>
   …
</iddf:data>
```


Typological Database System - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Back Forward Reload Stop Home
http://languagelink.let.uu.nl/tds/main.html#home[7]
Google
tds Typological Database System beta
TDS account
login or register
The query basket
contains 0 items
view clear
search by topic by datatype by language tutorial settings
Welcome to the Typological Database System
The Typological Database System (TDS) provides an online interface to multiple independently developed typological databases. It allows unified querying with the help of an integrated ontology.
Additional information can be found in the following places:
1. the tutorial tab explains the use of the TDS web interface;
2. the databases window lists the available databases and their details;
3. the frequently asked questions window lists the answers to common questions;
4. the about window contains further information about the project.
You can register yourself as a TDS user. This will allow your settings to be remembered for future sessions on this computer, and on any other computers you log on from. (You should therefore log out after using the TDS from a public computer). In the future there are more features planned which will be tied to a TDS user account, such as persistent storage of queries and notification when the results to a stored query change (because of additions to the TDS databases).
Support for "join" queries
It is now possible to perform "joins" in queries, that is, to execute a query that combines search terms about different independent data types. For example, the database treats languages, sentences and universal phoneme types as three separate data types in the database; you can now, for example, ask to see all sentences from languages that have ergative alignment (a language property).
2007 September 23rd
Instance browser
The instance browser provides the ability to navigate through all the information availabe on independent data types; you can, for example, browse all the information about the
menu
2004 - 2007 © LOT
mail comments/questions/etc. to the TDS project
the TDS server is reachable
Done
YSlow