UPPSALA
UNIVERSITET

XML as a preservation strategy

Experiences with the DiVA document format

Eva Müller, Uwe Klosa

Electronic Publishing Centre
Uppsala University Library, Sweden

UPPSALA UNIVERSITET

DiVA

Outline

- DiVA project and its objectives
- DiVA publishing system
- DiVA document format (DDF)
- Experiences with the DDF
- Conclusions and next steps

UPPSALA UNIVERSITET

DiVA

DiVA Project

- Start 2000; 2002 DiVA.1; 2004 DiVA.2
  Nine universities in three countries; number increasing
- Objectives:
  - Technical solutions & well functioning work flow supporting fulltext publishing, storage and dissemination of university research (theses, dissertations, working and research papers…)
  - Explore ways to ensure the future use and understanding of digital objects in the archive

UPPSALA UNIVERSITET

DiVA

… solutions focusing on

- Services
  - production
  - storage
  - preservation
  - retrieval
  - dissemination
- **Format** (metadata + stored documents)
- Work flows

UPPSALA UNIVERSITET

DiVA

Assumption that storage format is essential

Level of enabled **services** depends on granularity level of structure of the data stored within the system

Level of guarantees given for **future use and understanding** of the digital objects in the archive depends on the format used

UPPSALA UNIVERSITET

DiVA

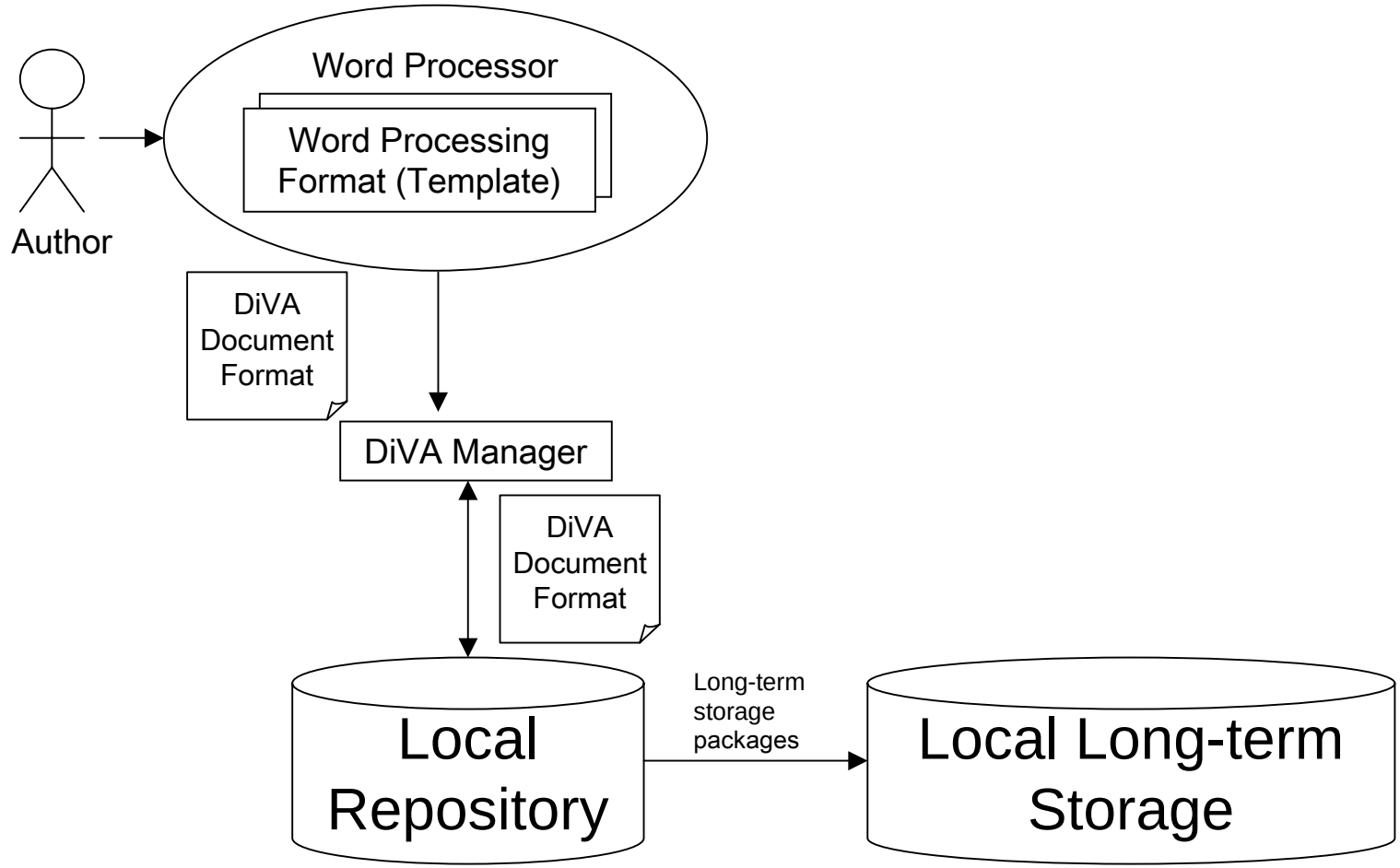
DiVA Publishing System

makes it possible to

- reuse and enhance the data directly from the source document originally created by authors, both for metadata and a digital master for electronic & printed versions
- assign a persistent identifier, store & checksum all files in a local archive
- send a copy to the national library archives and other interested parties

UPPSALA UNIVERSITET

DiVA



UPPSALA
UNIVERSITET

Implementation

- Java – XML technologies
- Currently an Oracle database used for indexing and searching
- Architecture: component-based design
  - Modularity and reusability of the components
  - Possibility to seamlessly replace modules with improved implementations of the component

UPPSALA UNIVERSITET

DiVA

DiVA Document Format (DDF)

- Internal format developed for, but not limited to, academic publications
- Version 1.0 (defined by an XML Schema)
  - http://publications.uu.se/schema/ddf/
  - Component based
  - Extensible
  - Administrative metadata elements are combined with descriptive elements
  - Elements conforming to DocBook DTD are used for the content part of the document

UPPSALA UNIVERSITET

DiVA

Why a customized format?

DiVA document – the result of a practical approach demanding

- self-description
- clear structure
- support for export to other formats/schemas
- compatibility with a number of metadata formats/schemas
- easy reuse of data

UPPSALA UNIVERSITET

DiVA

Why XML?

- Open and established notation
- Support for international character sets (UNICODE)
- A simple and human readable text format

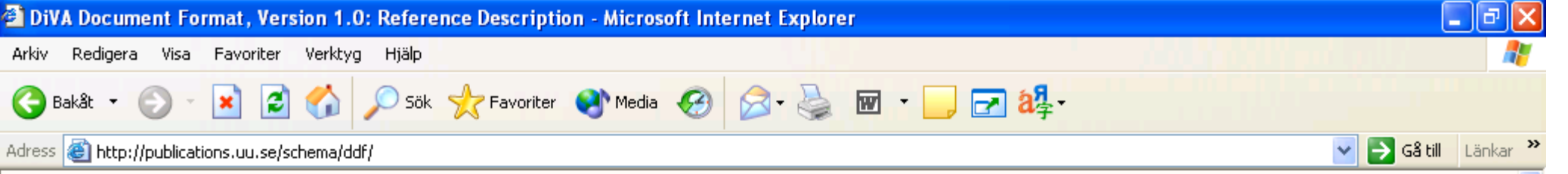
… characteristics facilitate data migration and the documents are likely to have longevity

UPPSALA
UNIVERSITET

Why XML Schema rather than DTD?

"XML schema provides means for defining the structure, content and semantics of XML documents"

- It's written in XML
- It supports data types, self-defined data types and namespaces
  - validation, restriction definition, data format definition …



DiVA Document Format

Version 1.0: Reference Description (Draft)

Table of Contents

http://publications.uu.se/schema/ddf/

UPPSALA UNIVERSITET

The global structure of the DiVA document

Metadata description of publication, which may contain fulltext document

- Root element documents to allow many documents to be included in a single file
- Each individual document is described within document element
- If the fulltext is included it appears within the contents element

UPPSALA UNIVERSITET

DiVA

…global structure of the DiVA XML document

```
<documents>
  <date type="creation" timezone="UTC+1">
    <year>2004</year>
    <month>01</month>
    <day>27</day>
  </date>
  <time type="creation"
   timezone="UTC+1">14:28</time>
  <document>
     ...the metadata ...
    <contents>...the fulltext contents...</contents>
  </document>
</documents>
```

UPPSALA UNIVERSITET

DiVA

Metadata

http://publications.uu.se/schema/ddf/divametadata.html#DocumentStructure

- Common elements

  [e.g. properties, identifiers, specifics, languages, creators, contributors, titles, abstracts, contents, note ..]

- Manifestations - container for one or more manifestation elements that contain metadata about a particular format of the document

  [e.g. properties, date, time, edition, publisher, distributors, archivers …]

- Mappings

  [DC/RDF, MARCXML, METS, MODS, TEI Header, Endnote, MARC21…]

UPPSALA
UNIVERSITET

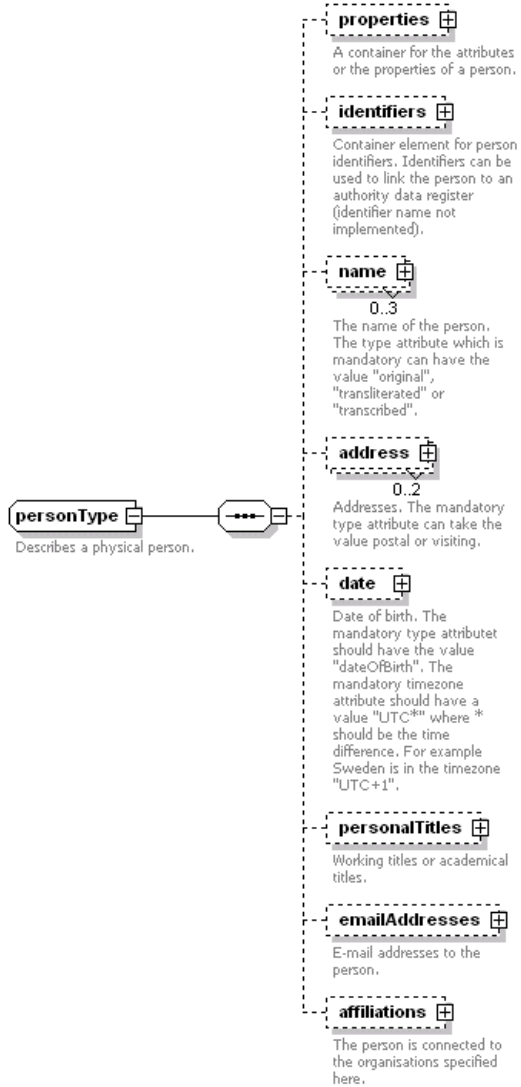
Components

For example

- addressType
- personType
- organisationType

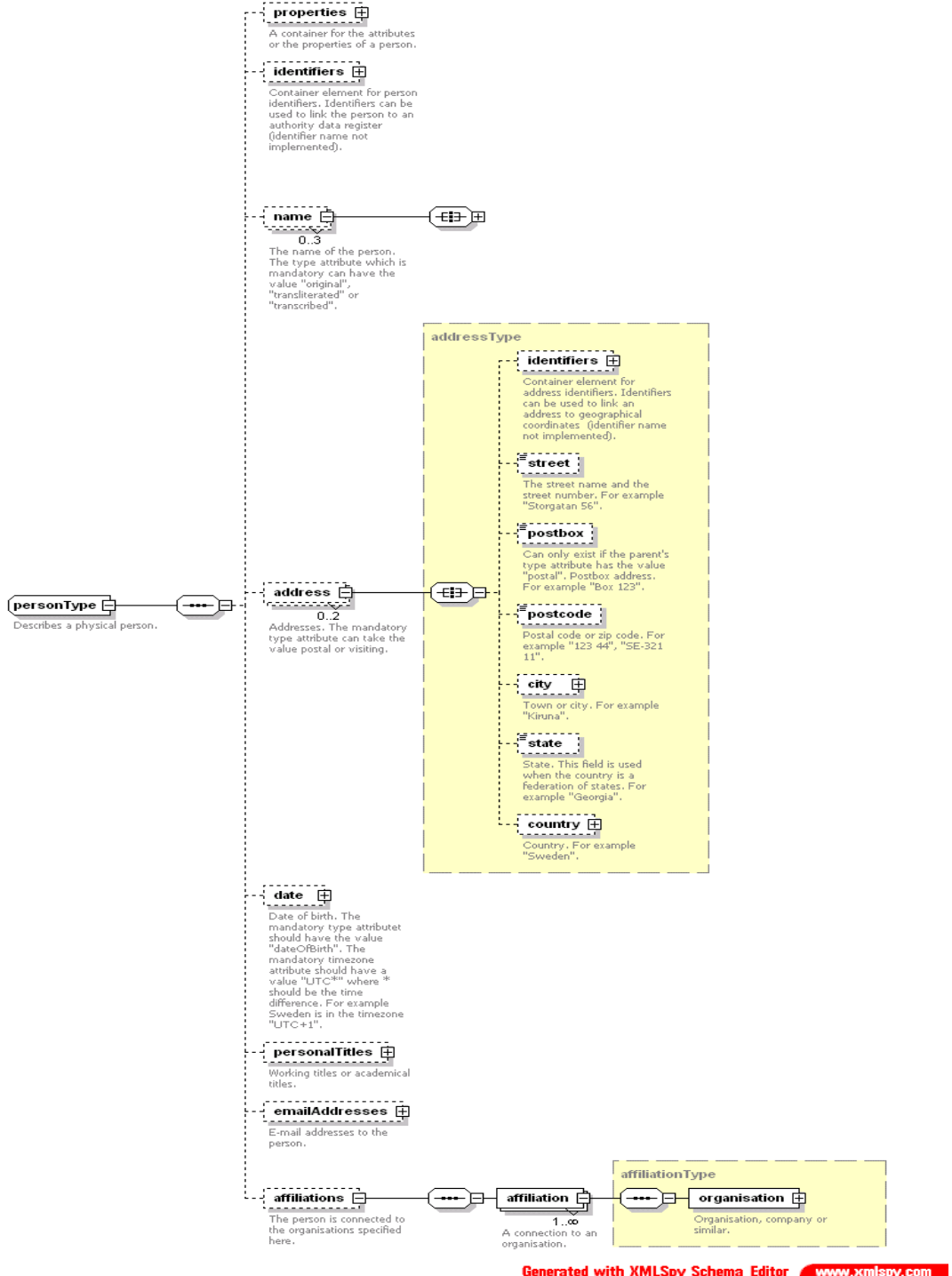
UPPSALA UNIVERSITET

DiVA



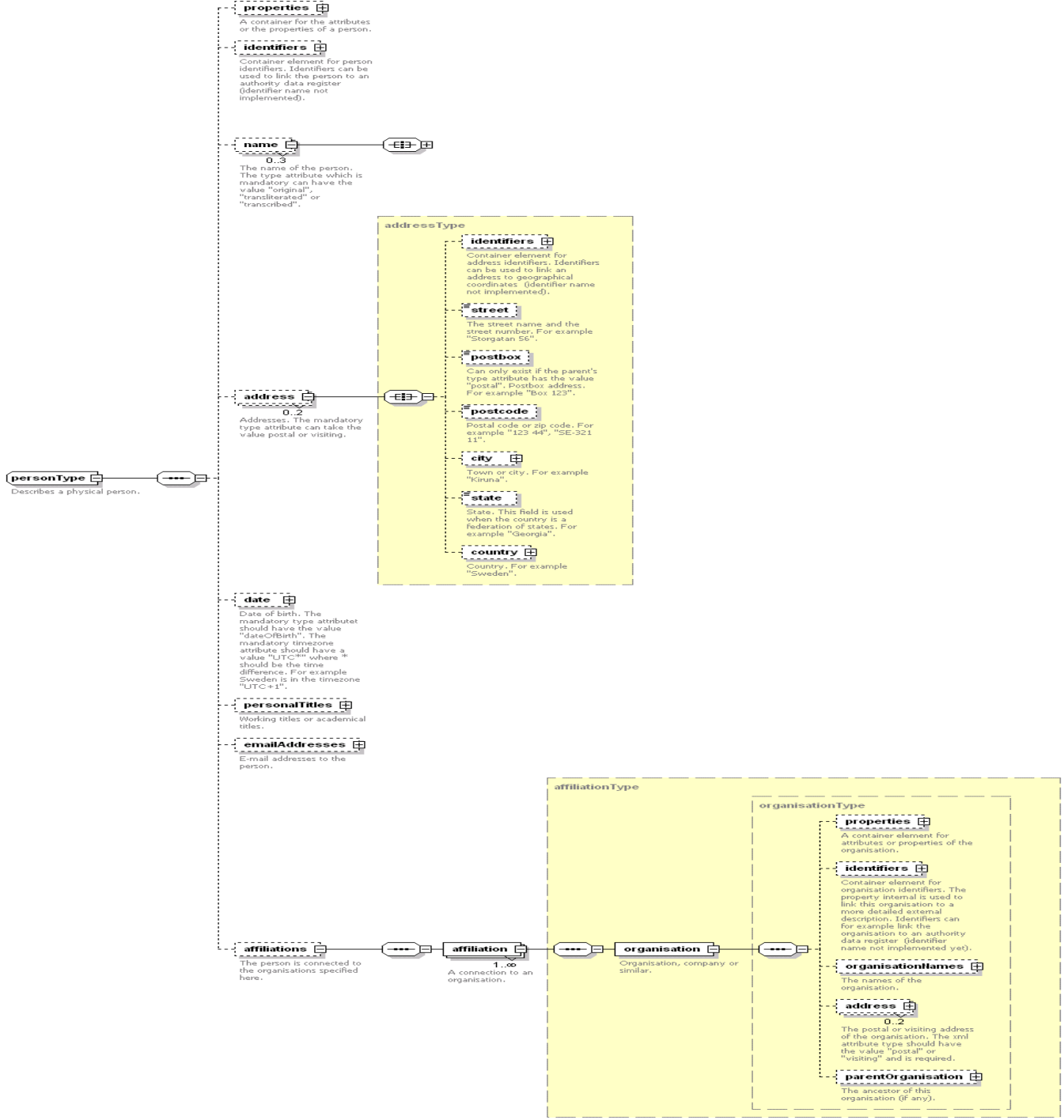
UPPSALA UNIVERSITET

DiVA



UPPSALA UNIVERSITET

DiVA

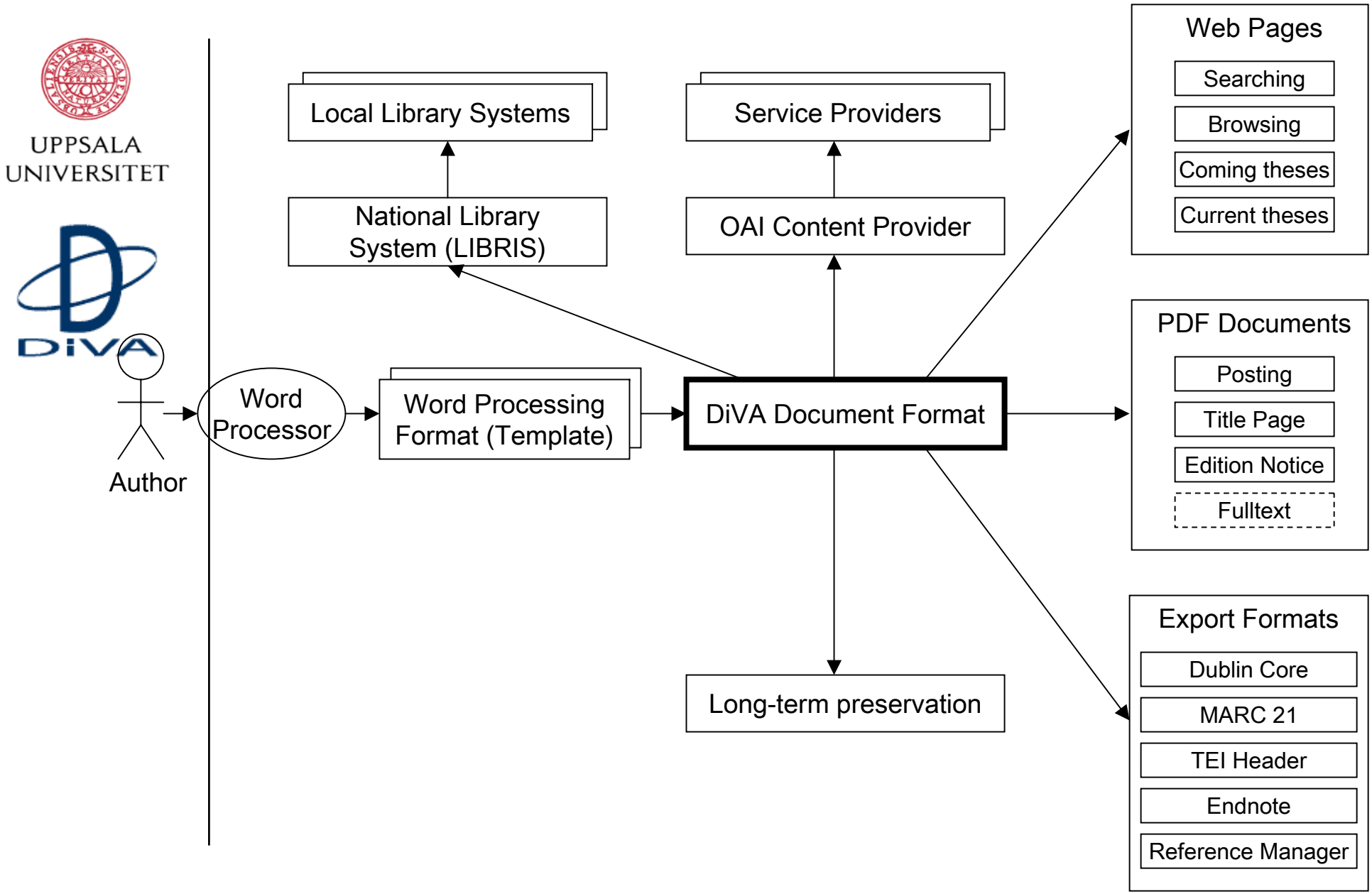


UPPSALA UNIVERSITET

DiVA

DDF in the DiVA system

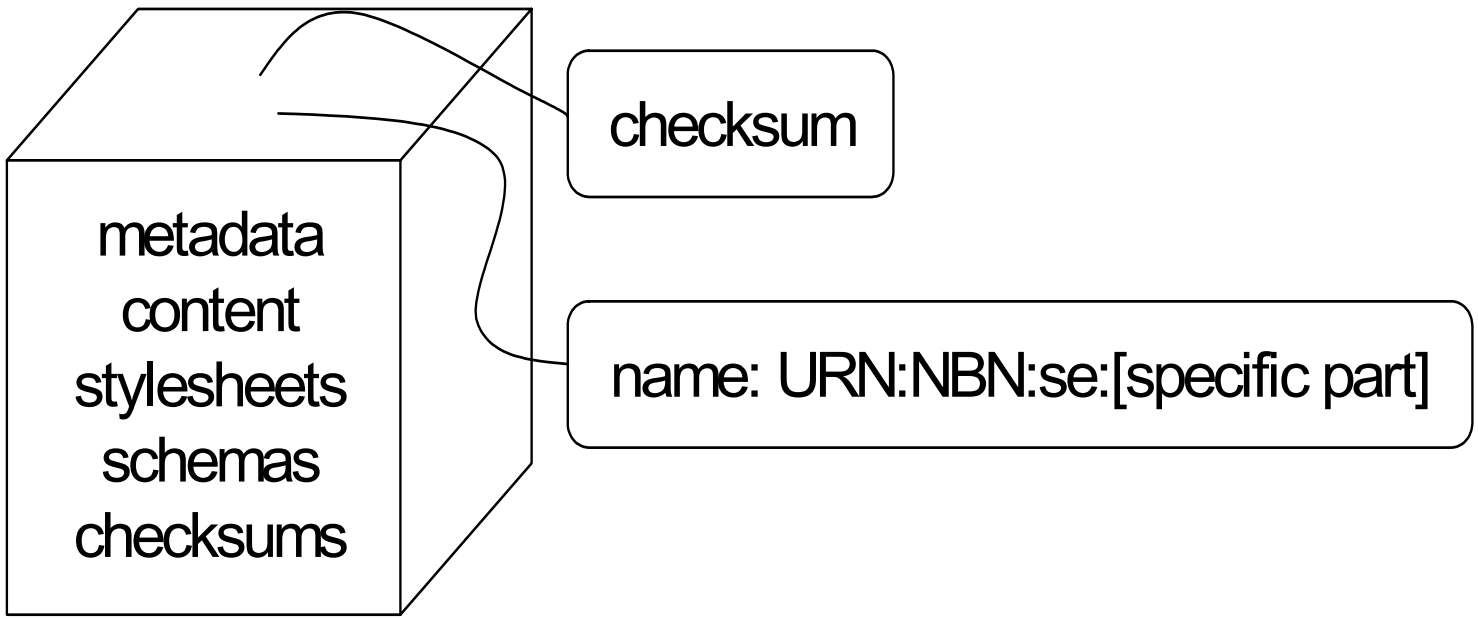
- Subsystem interface
- Source for other formats (DocBook, TEI, Marc 21, DC etc.)
- Long-term preservation format


UPPSALA UNIVERSITET
DiVA
Author
Word Processor
Word Processing Format (Template)
Local Library Systems
National Library System (LIBRIS)
Service Providers
OAI Content Provider
DiVA Document Format
Long-term preservation
Web Pages
Searching
Browsing
Coming theses
Current theses
PDF Documents
Posting
Title Page
Edition Notice
Fulltext
Export Formats
Dublin Core
MARC 21
TEI Header
Endnote
Reference Manager

UPPSALA UNIVERSITET

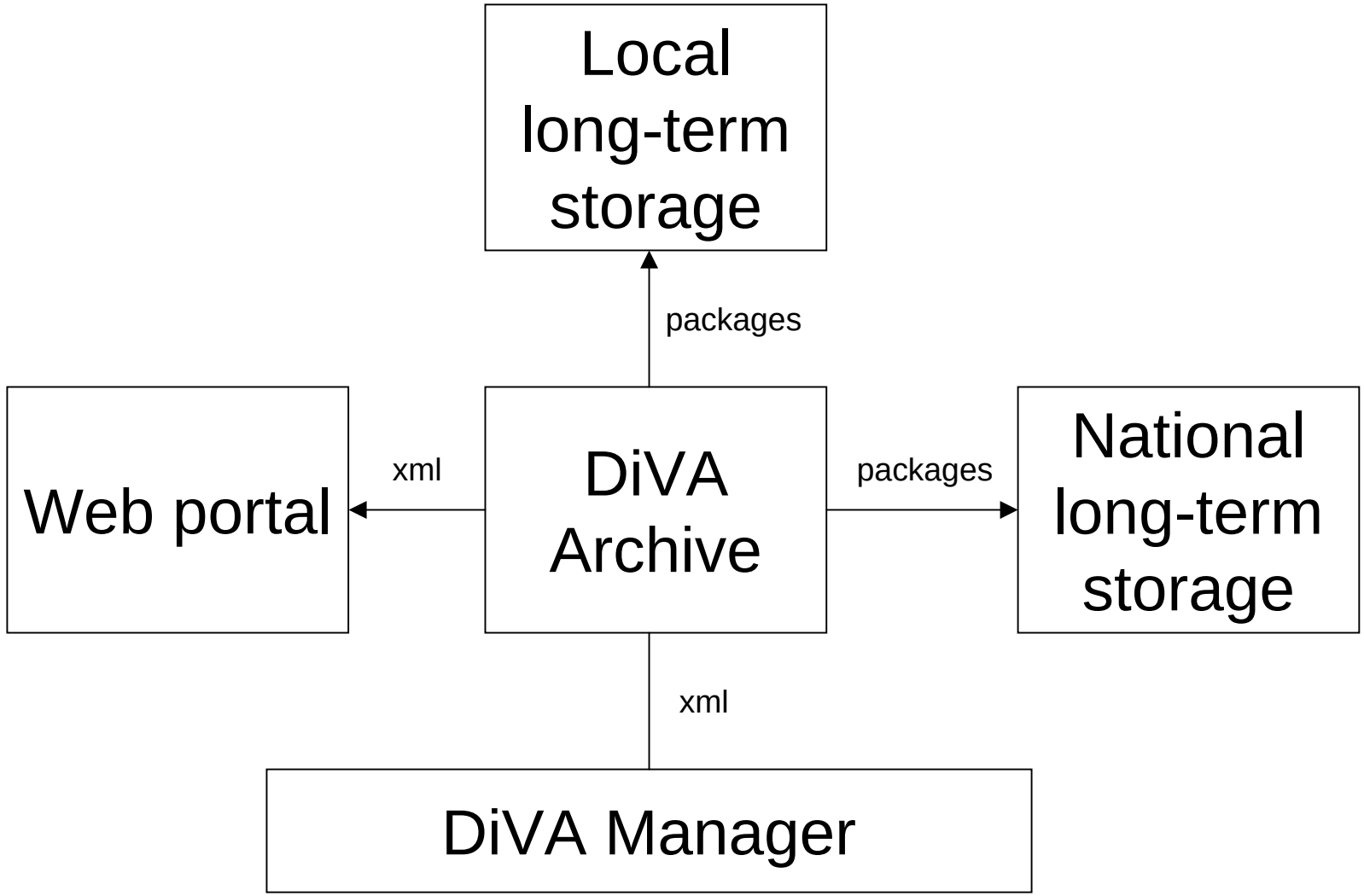
DiVA

Archiving Package



UPPSALA UNIVERSITET

DiVA



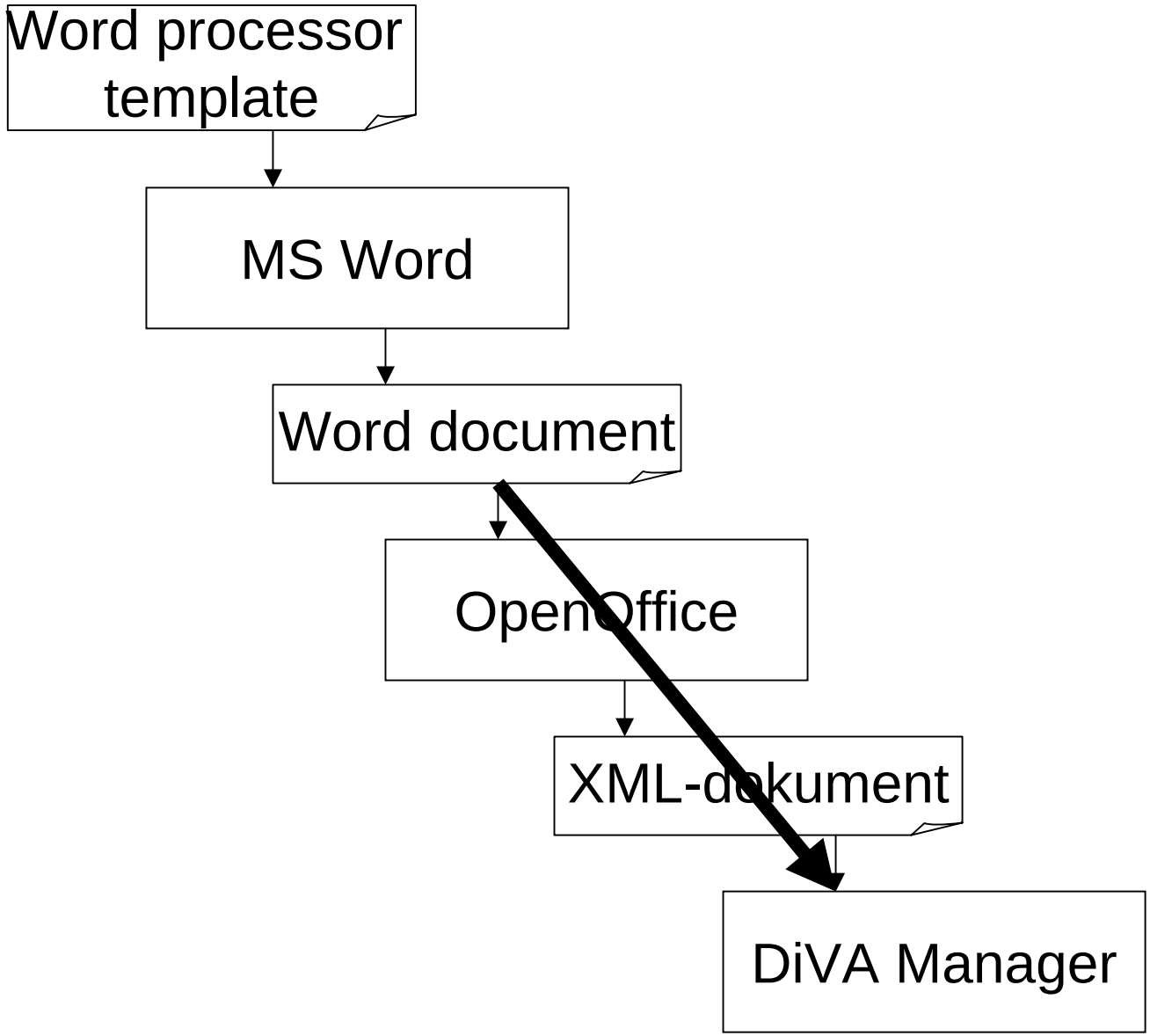
UPPSALA UNIVERSITET

DiVA

How is the format produced?

- Automated work flow based on delivery of information in templates
- Tools for conversion from Word processing templates (MS Word/Open Office – XML)
- Work in progress – MathML + images
- Demo – http://

Import


Word processor template
MS Word
Word document
OpenOffice
XML-dokument
DiVA Manager

UPPSALA UNIVERSITET

DiVA

Next steps

- Revisions and extensions of DDF
  - Multiple file manifestations
  - Rights metadata
  - Extended preservation metadata
  - Relations to other resources
- DiVA.2 scheduled January 2005
  - open for comments

UPPSALA UNIVERSITET

DiVA

More information

- **The DiVA Project - Development of an Electronic Publishing System** [English]
(D-Lib Magazine, (9)2003:11)
http://www.dlib.org/dlib/november03/muller/11muller.html
- **Archiving Workflow Between a Local Repository and the National Archive** [English]
(2003-08-18: ECDL 2003, Web Archives, Workshop)
http://publications.uu.se/epcentre/conferences/ecdl2003/archiving_ECDL_2003.pdf
- **Using XML for Long-term Preservation : Experiences from the DiVA Project** [English]
(2003-05-22: ETD 2003: Next Steps - Electronic Theses and Dissertations Worldwide, Berlin)
http://publications.uu.se/etd2003/papers/LongTermPreservation.pdf
- DiVA portal – http://www.diva-portal.org/