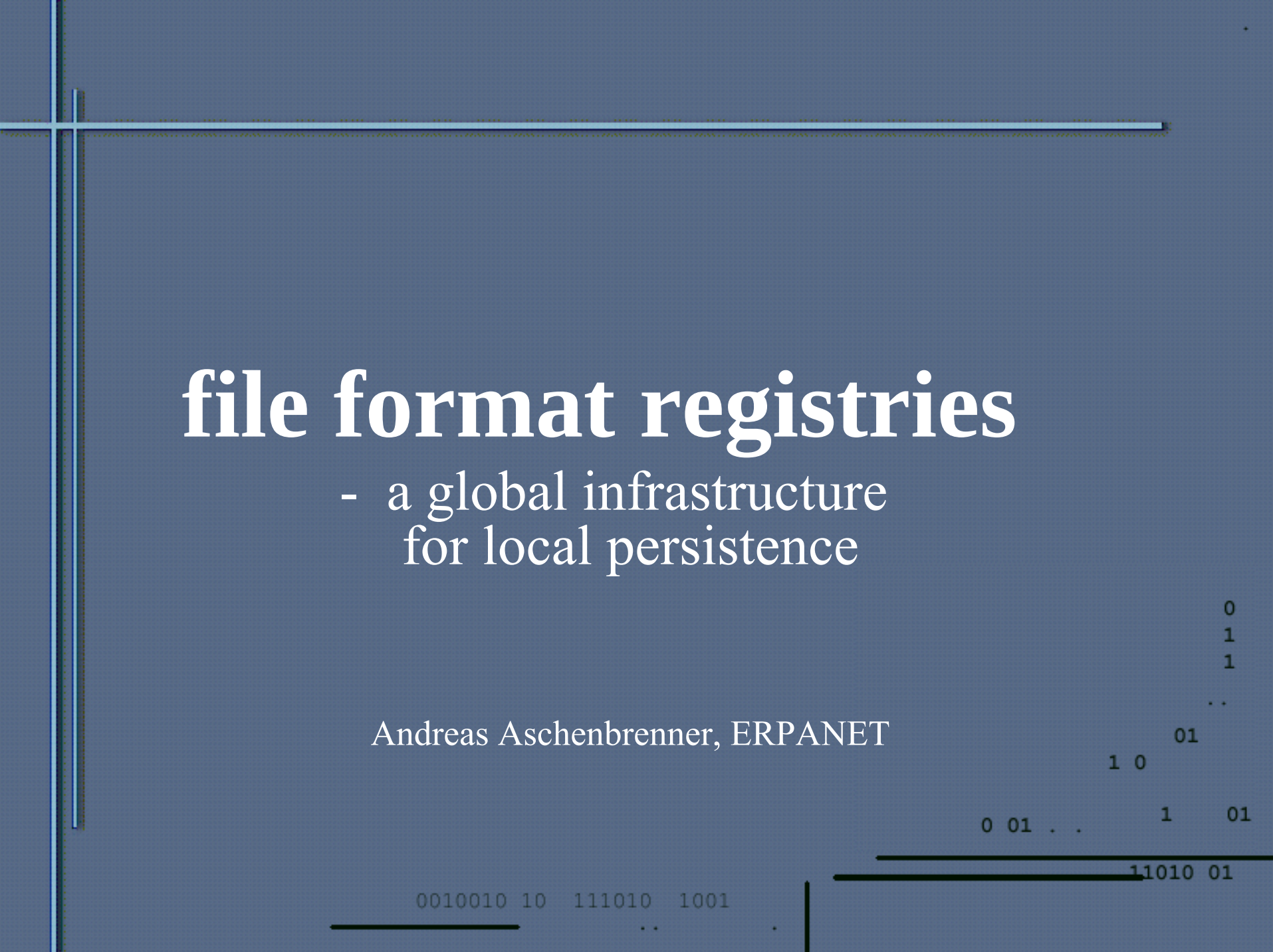
file format registries

- a global infrastructure for local persistence

Andreas Aschenbrenner, ERPANET



overview

- motivation
- registry features
- PRONOM
- Global Digital Format Registry

the shared need

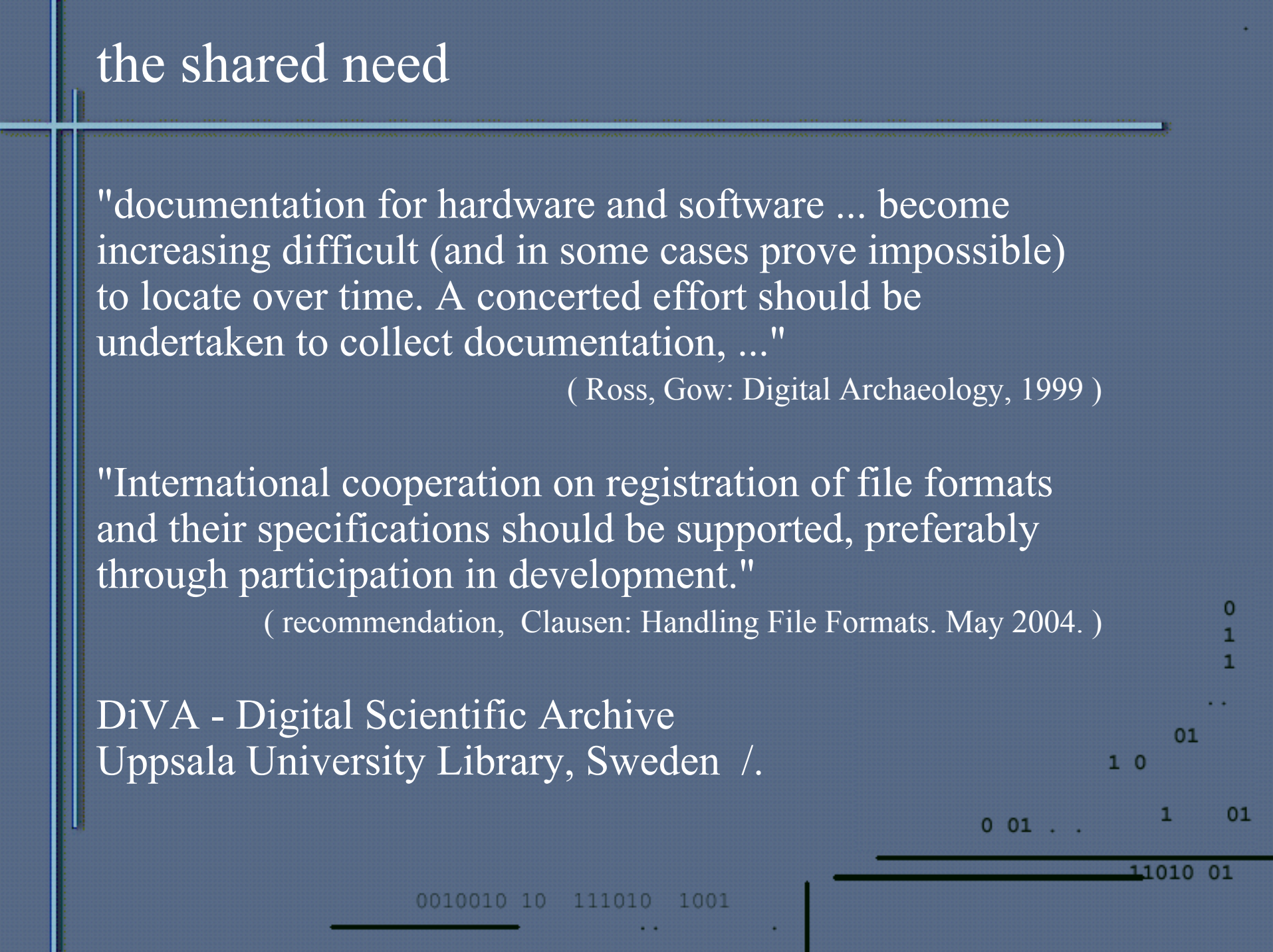
"documentation for hardware and software ... become increasing difficult (and in some cases prove impossible) to locate over time. A concerted effort should be undertaken to collect documentation, ..."

(Ross, Gow: Digital Archaeology, 1999)

"International cooperation on registration of file formats and their specifications should be supported, preferably through participation in development."

(recommendation, Clausen: Handling File Formats. May 2004.)

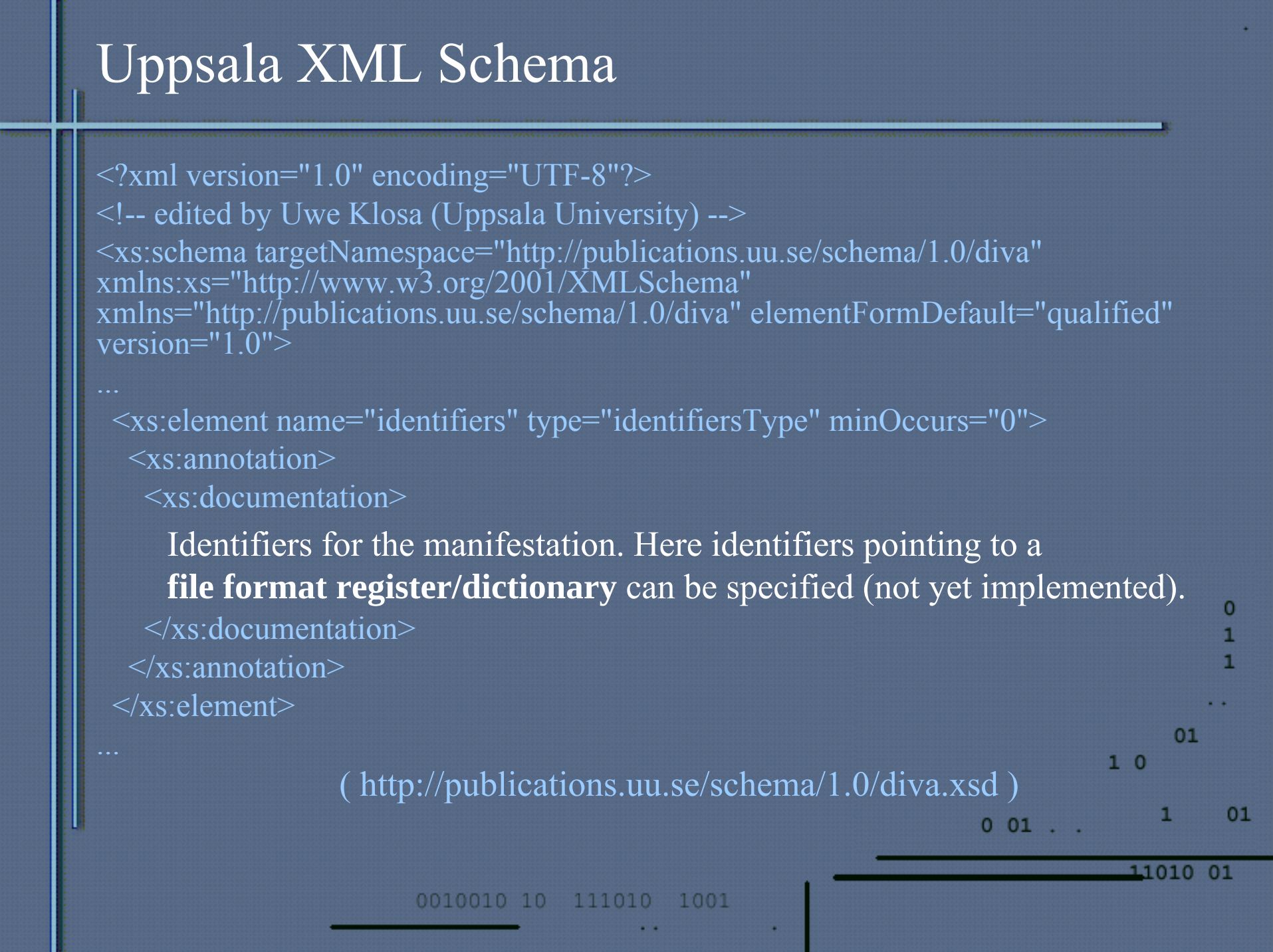
DiVA - Digital Scientific Archive
Uppsala University Library, Sweden /.



Uppsala XML Schema

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- edited by Uwe Klosa (Uppsala University) -->
<xs:schema targetNamespace="http://publications.uu.se/schema/1.0/diva"
xmlns:xs="http://www.w3.org/2001/XMLSchema"
xmlns="http://publications.uu.se/schema/1.0/diva" elementFormDefault="qualified"
version="1.0">
...
  <xs:element name="identifiers" type="identifiersType" minOccurs="0">
    <xs:annotation>
      <xs:documentation>
        Identifiers for the manifestation. Here identifiers pointing to a
        file format register/dictionary can be specified (not yet implemented).
      </xs:documentation>
    </xs:annotation>
  </xs:element>
...
```

(http://publications.uu.se/schema/1.0/diva.xsd)



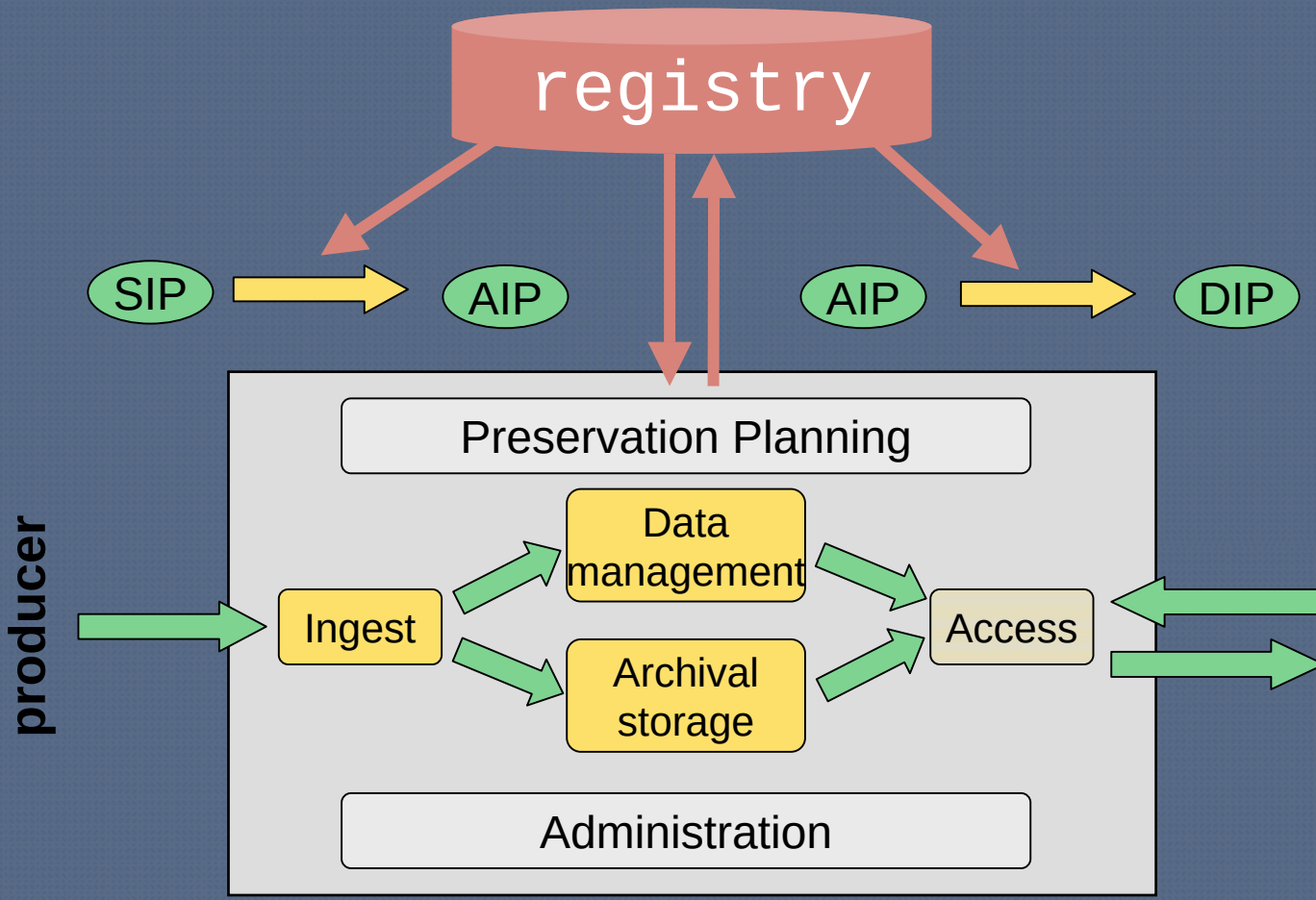
representation networks

Representation Information: The information that maps a Data Object into more meaningful concepts.

Representation Network: The set of Representation Information that fully describes the meaning of a Data Object.

(OAIS Model)
(→ Cedars 1999)

OAIS model


registry
SIP
AIP
AIP
DIP
Preservation Planning
producer
Data management
Ingest
Access
consumer
Archival storage
Administration

registry use cases

Identification

- "I have a digital object; what format is it?"

Validation

- "I have an object purportedly of format *F*; is it?"

Transformation

- "I have an object of format *F*, but need *G*; how can I produce it?"

Characterization

- "I have an object of format *F*; what are its features?"

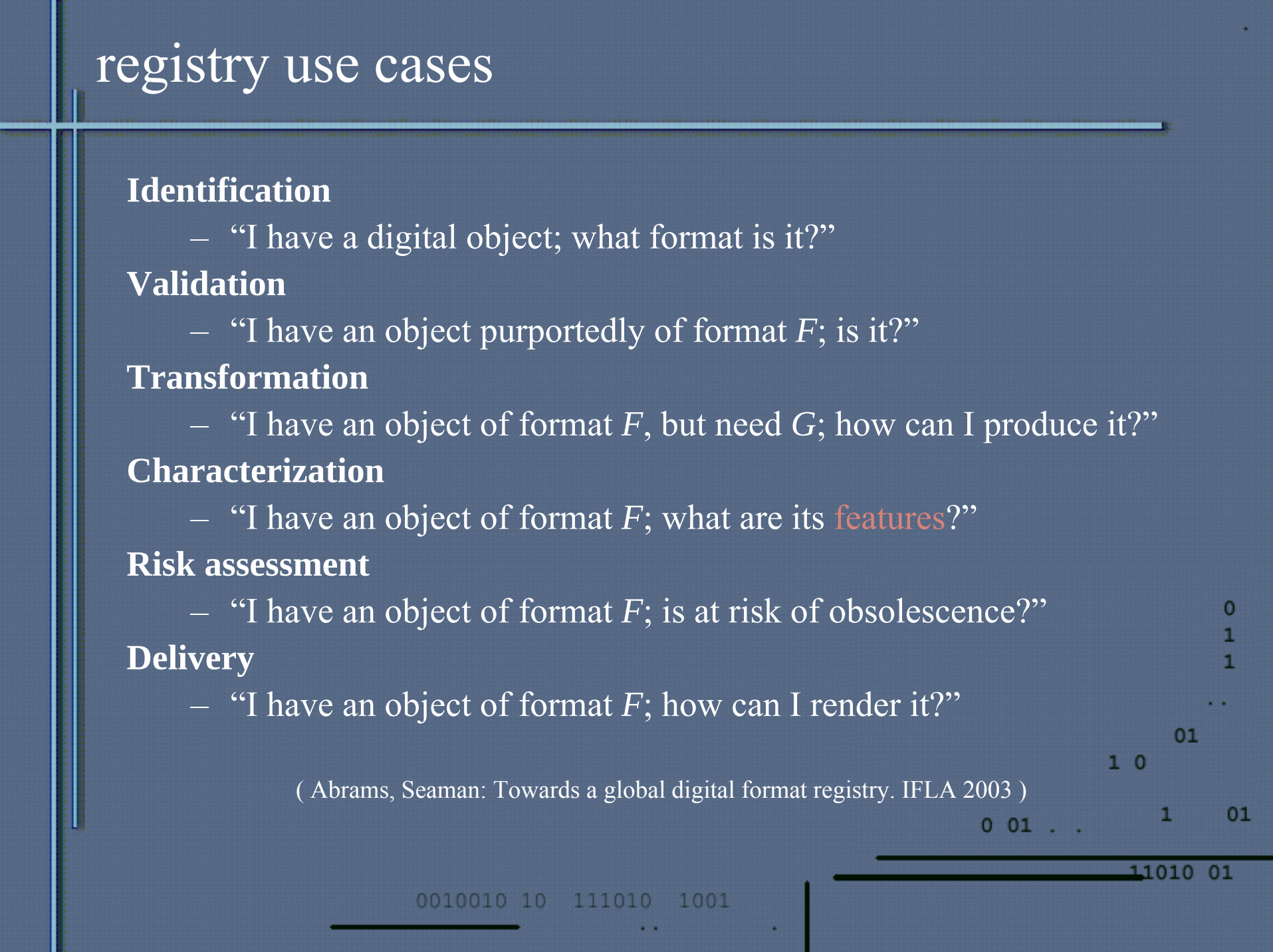
Risk assessment

- "I have an object of format *F*; is at risk of obsolescence?"

Delivery

- "I have an object of format *F*; how can I render it?"

(Abrams, Seaman: Towards a global digital format registry. IFLA 2003)



PRONOM

* UK National Archives, 2001

  "PRONOM is a resource for anyone requiring impartial and definitive technical information about the file formats used to store electronic records, and the software products that are required to create, render, or migrate these formats."

→ operative since March 2002

→ opened web access January 2004

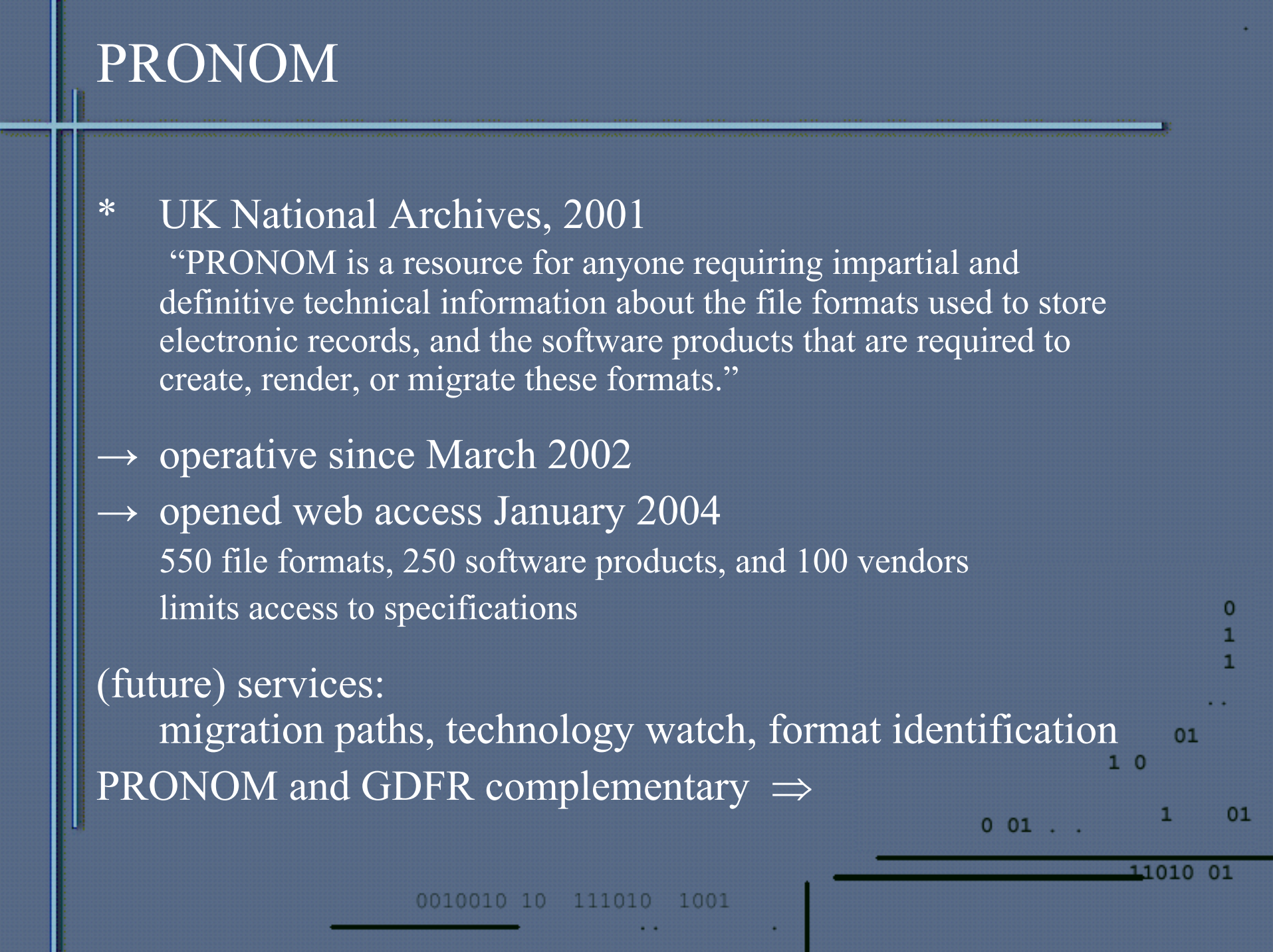
  550 file formats, 250 software products, and 100 vendors

  limits access to specifications

(future) services:

  migration paths, technology watch, format identification

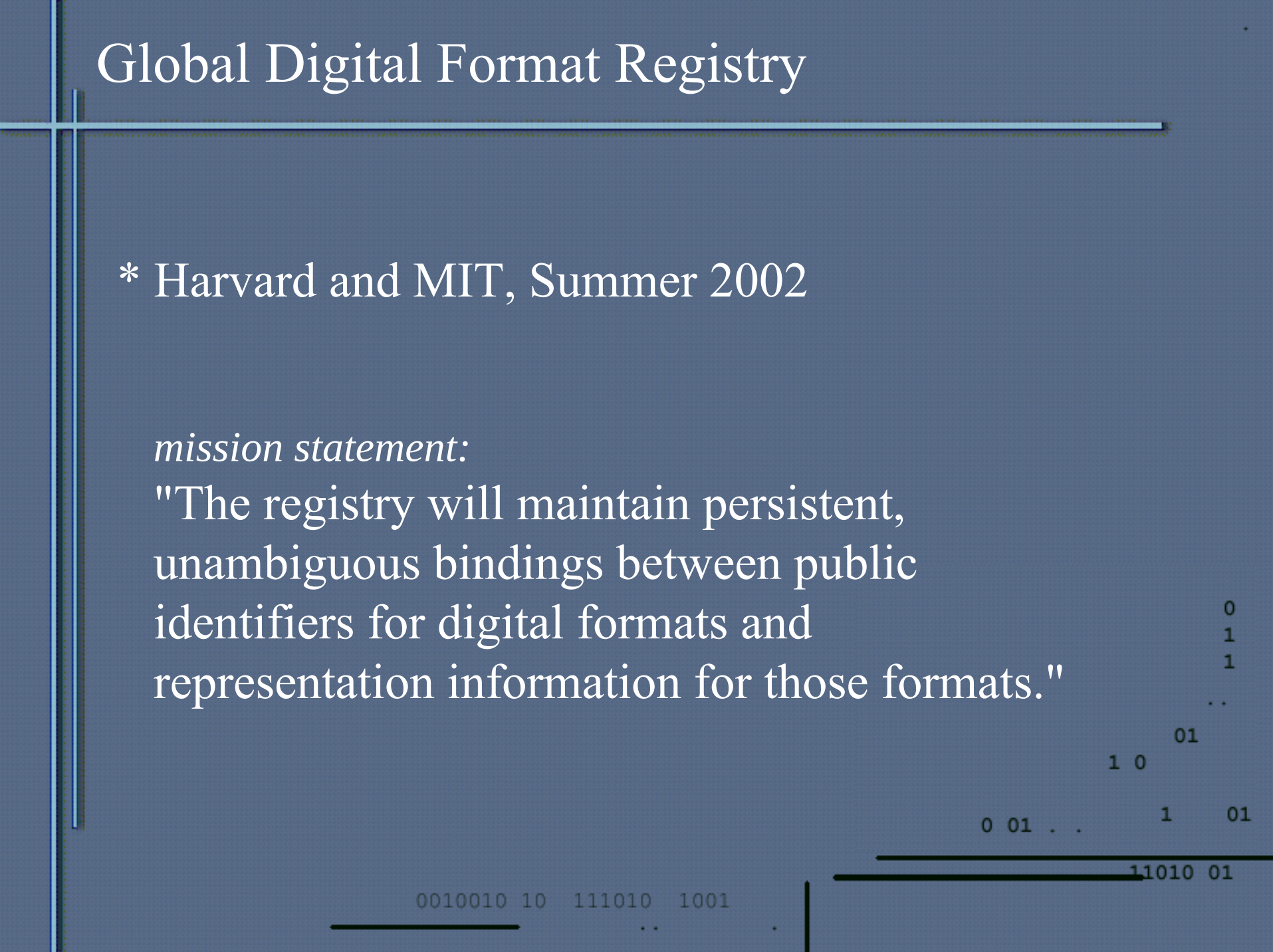
PRONOM and GDFR complementary ⇒



Global Digital Format Registry

* Harvard and MIT, Summer 2002

mission statement:
"The registry will maintain persistent, unambiguous bindings between public identifiers for digital formats and representation information for those formats."



Global Digital Format Registry

Ad-Hoc Committee

Bibliothèque Nationale, France
British Library
California Digital Library
Digital Library Federation
Harvard University
IETF
Internet Architecture Board
JISC
JSTOR
Library of Congress
MIT
NARA
National Archives of Canada
National Archives, UK
New York University
NIST
OCLC
University of Pennsylvania
RLG
Stanford University

Global Digital Format Registry

design and implementation phase

funded through grants

developed data model

- descriptive: identifier, ontology, format relationships,
- characterisation: specification document, signature

operational phase

must be ***trustworthy*** and ***sustainable***

how to populate and maintain registry?

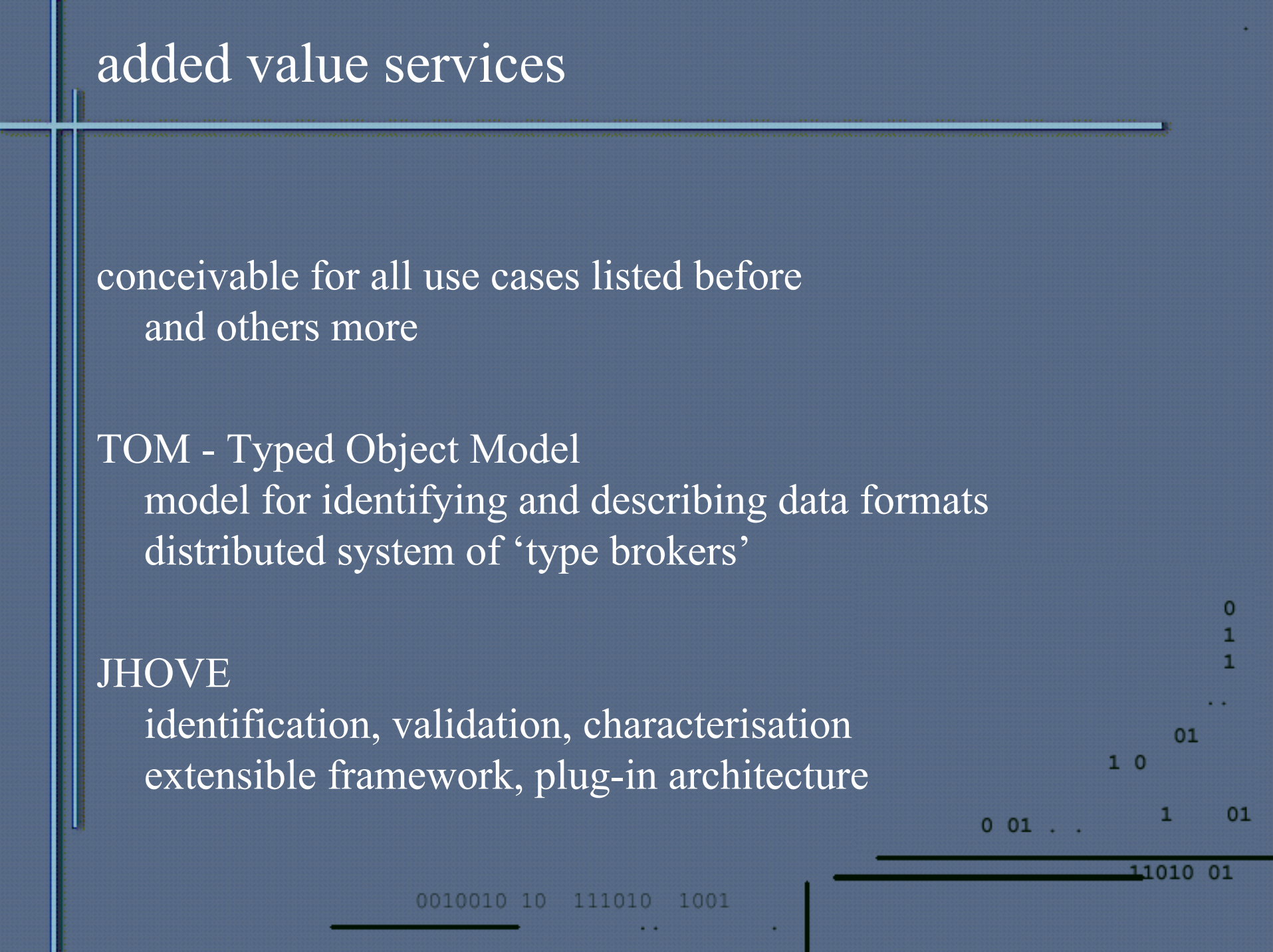
centralised vs distributed registry?

added value services

conceivable for all use cases listed before
- and others more

TOM - Typed Object Model
- model for identifying and describing data formats
- distributed system of ‘type brokers’

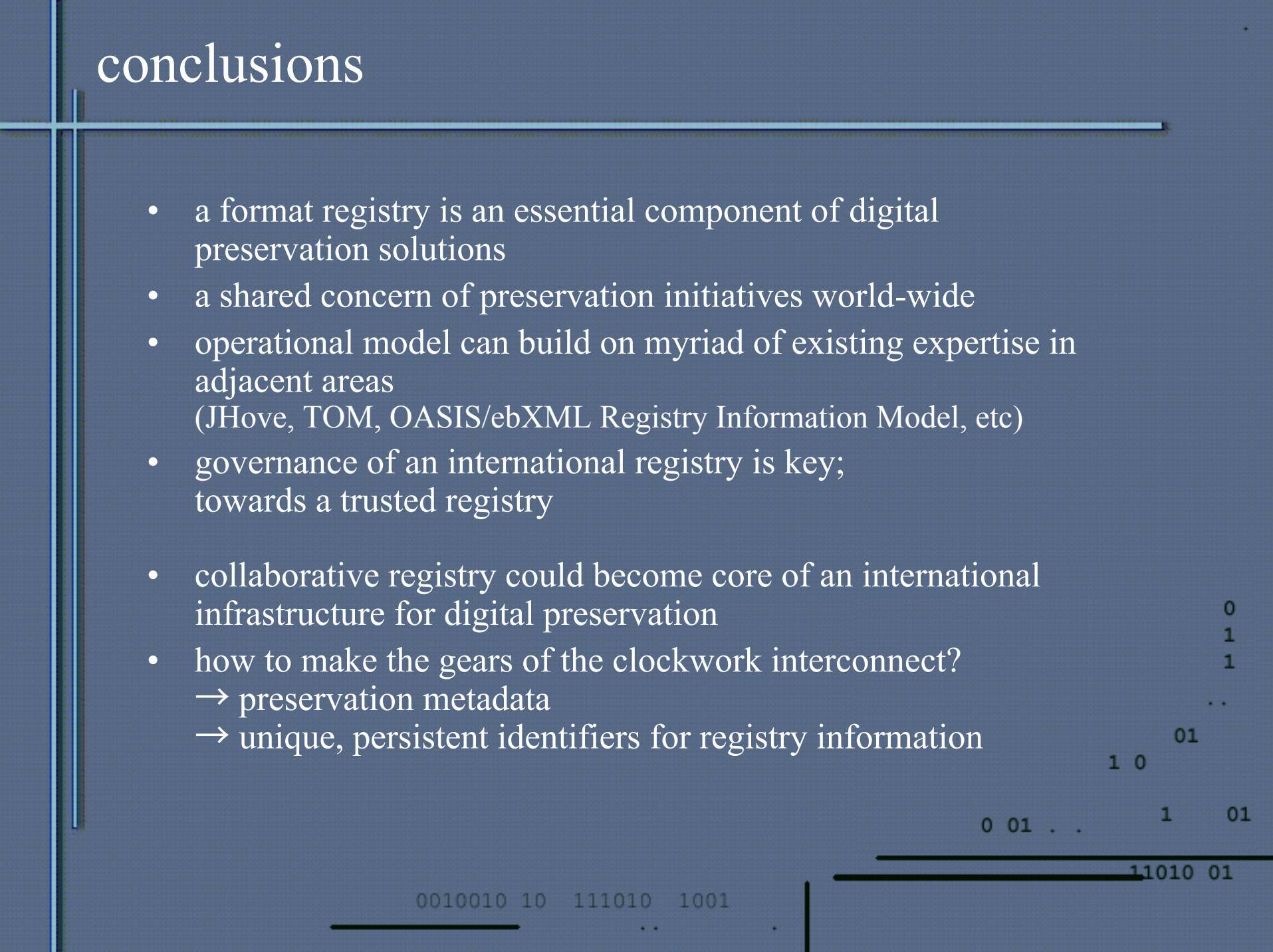
JHOVE
- identification, validation, characterisation
- extensible framework, plug-in architecture



conclusions

- a format registry is an essential component of digital preservation solutions
- a shared concern of preservation initiatives world-wide
- operational model can build on myriad of existing expertise in adjacent areas
  (JHove, TOM, OASIS/ebXML Registry Information Model, etc)
- governance of an international registry is key;
  towards a trusted registry

- collaborative registry could become core of an international infrastructure for digital preservation
- how to make the gears of the clockwork interconnect?
  → preservation metadata
  → unique, persistent identifiers for registry information



further reading

* Global Digital Format Registry (GDFR)
  http://hul.harvard.edu/gdfr/
* PRONOM, UK National Archives:
  http://www.records.pro.gov.uk/pronom/
* University of Pennsylvania Library, John Mark Ockerbloom:
  TOM - Typed Object Model: http://tom.library.upenn.edu/
  FRED - Format REgistry Demo.: http://tom.library.upenn.edu/fred/
* JHOVE: http://hul.harvard.edu/jhove/
* Abrams, Seaman: Towards a global digital format registry.
  69th IFLA 2003.
  http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf
* Stephen L. Abrams: Global Digital Format Registry. Presentation at RLG/CIMI “Ready to Wear” New York, May 12-13, 2003.
  http://www.rlg.org/events/metadata2003/abrams.ppt
* Representation and Rendering Project: File Format Report. 2003.
  http://www.leeds.ac.uk/reprend/