

# Chinese-European Workshop on Digital Preservation

Beijing (China), July 14 – 16, 2004

## Why Digital Preservation? Needs and Challenges



**Dr. Heike Neuroth**

**Research & Development**

**Göttingen State and University Library (SUB)**

**[neuroth@mail.sub.uni-goettingen.de](mailto:neuroth@mail.sub.uni-goettingen.de)**





## ToC

- Long-term Preservation
  - Relevance for (scientific) libraries
  - Heterogeneity
- Model (?)
- Presentations
- Outlook



## Long-term Presevation

- New forms of production, publication and distribution of scientific information
- Rapid change of technology, different/various ways of preservation
- Trusted repositories needed
  - well defined criteria (policy etc.)
  - certification? (e.g. DINI)
- Guarantee of trusted long-term preservation
- IPR, Digital Rights Management
- ...



...

Digital Preservation consists of processes to ensure that digital resources remain **accessible**, **usable** and **understandable** in the future.

→ To ensure that future software and hardware tools will generate an authentic and integral representation of the object



...

What is meant by „**long term**“?

→ Definition by Ute Schwens / Hans Liegmann:

- In terms of preserving digital resources, ‚long-term‘ does not mean issuing a guarantee for five or fifty years, rather the **responsible development of strategies** which can cope with the constant changes brought about by the information market.



## How much information?

- UC Berkeley's School of Information Management and Systems: How much Information?
  - Analyse of year 2002 to estimate the yearly increase of new (digital and analog) information
    - physical/storage media: print, film, magnetic, optical
    - information flows: telephone, radio, TV, Internet

[October 2003]



- **5 exabytes** of new information in 2002  
(= 0,5 Mio x Print-Collection of LoC)
    - storage media – magnetic:  
**80%** increase, of which **92 %** are stored on hard disc!
    - information flow – Internet, only surface (!!):  
**170 terabytes** of information (= 17 x LoC)
- **30 %** Increase of digital information per year  
(so far)

**Table 1.2: Worldwide production of original information, if stored digitally, in terabytes circa 2002. Upper estimates assume information is digitally scanned, lower estimates assume digital content has been compressed.**

Storage Medium	2002 Terabytes Upper Estimate	2002 Terabytes Lower Estimate	1999-2000 Upper Estimate	1999-2000 Lower Estimate	% Change Upper Estimates
Paper	1,634	327	1,200	240	36%
Film	420,254	76,69	431,690	58,209	-3%
Magnetic	4,999,230	3,416,230	2,779,760	2,073,760	80%
Optical	103	51	81	29	28%
<b>TOTAL:</b>	<b>5,421,221</b>	<b>3,416,281</b>	<b>3,212,731</b>	<b>2,132,238</b>	<b>69%</b>

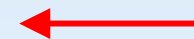
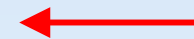
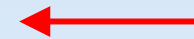
Source: *How much information 2003*



**Table 1.14: World Distribution of Internet Users (in millions)**

Africa	6.31
Asia Pacific	187.24
Europe	190.91
Middle East	5.12
Canada and USA	182.67
Latin America	33.35

Source: *Nielsen / NetRatings via CyberAtlas*





**Table 1.13: The size of the Internet in terabytes.**

<b>Medium</b>	<b>2002 Terabytes</b>
Surface Web	167
Deep Web	91,850
Email (originals)	440,606
Instant messaging	274
<b>TOTAL</b>	<b>532,897</b>

*Source: How much information 2003*

**10 TB = Print-Collection of LoC**



## Relevance

- (Scientific) libraries have to provide long-term access to scientific resources
  - regardless of the format
  - regardless of the document type
  - across all disciplines
- In Germany: DFG (German Research Foundation)
  - SSG libraries have a mandate to provide access to subject-specific scientific objects and to preserve them



## Heterogeneity: Document-Types

- Journals and monographs at SUB (ca. 1.5 Mio)
  - retrodigitized material (e.g. Springer)
  - genuin digital material
  - different formats: PDF, TEX, TIFF, etc.
- Web-Documents, Web-Server
- Preprint-Server, Theses, e-Proceedings, etc.
- Primary data
- CD's
- ...



## Heterogeneity: Format-Types

- Depends on subject, e.g.
  - Mathematics (TEX, PS, ...)
  - Geography (GIS)
  - ...
- Multimedia, e.g.
  - Animated WWW pages
  - Interactive objects in e-Learning
- Different versions in e.g. PDF, TEX, ...
- ...
- **Presentation Format + Preservation Format**



## Heterogeneity: General

- Metadata formats (Dublin Core, MODS, ..)
- Exchange formats (XML, METS, XML/RDF, ...)
- Controlled vocabulary systems (Ontologies, Taxonomies, ...)
- Architecture, Protocols
- ...

**Standardisation!**

**Interoperability!**



## Strategy

- Preservation policy
- Cooperation: international/national
- Cooperation: cross-domain
- Redundance of digital repositories explicitly desired
- Cooperative management/administration of distributed digital archives/repositories

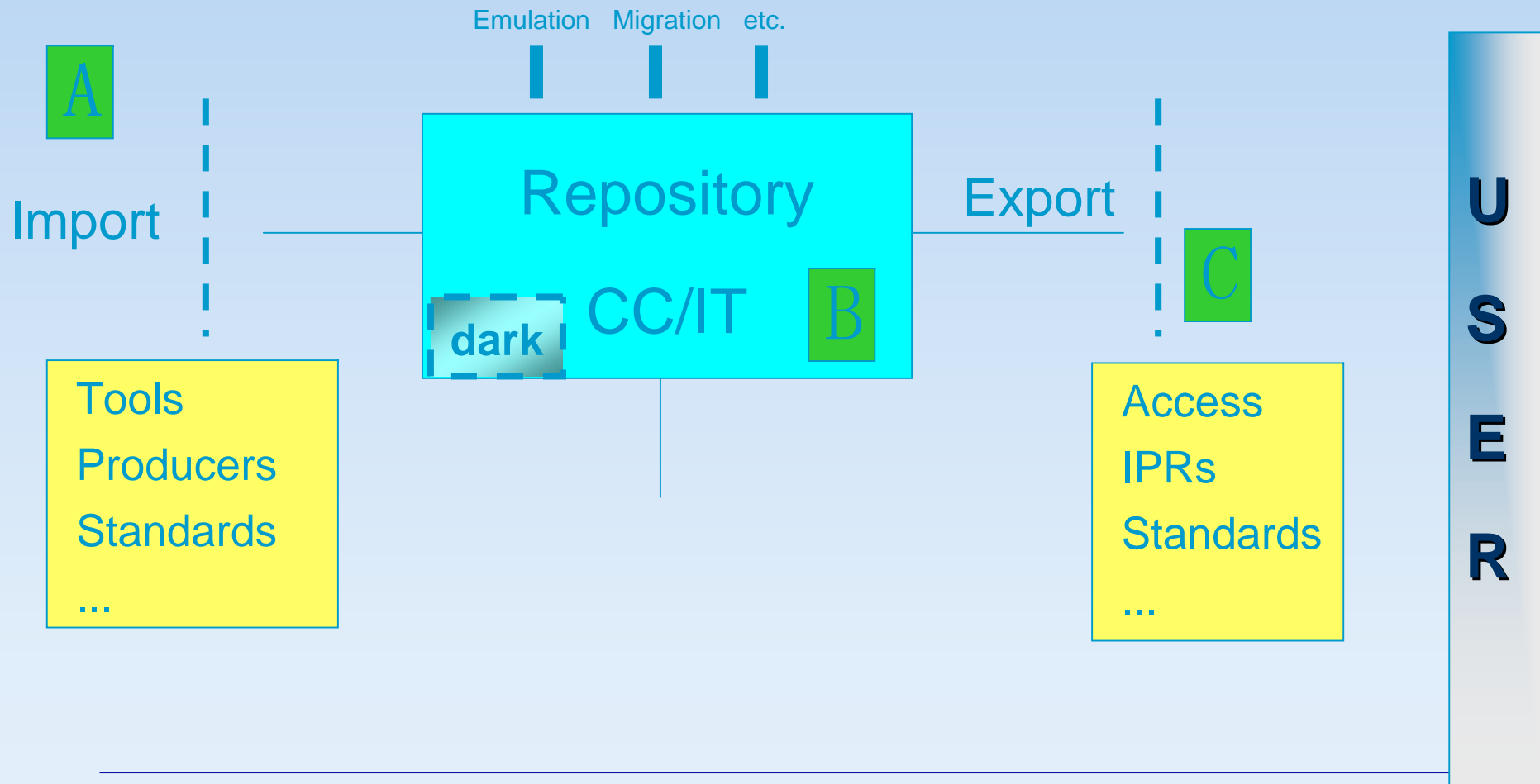


...

- Coordinated cooperation needed between:
  - Producer of digital objects (e.g. scientist)
  - Provider (e.g. library)
  - Distributor (e.g. publisher, hosts of db)
  
- International standards (e.g. DC, OAI, OAIS, METS), interoperability



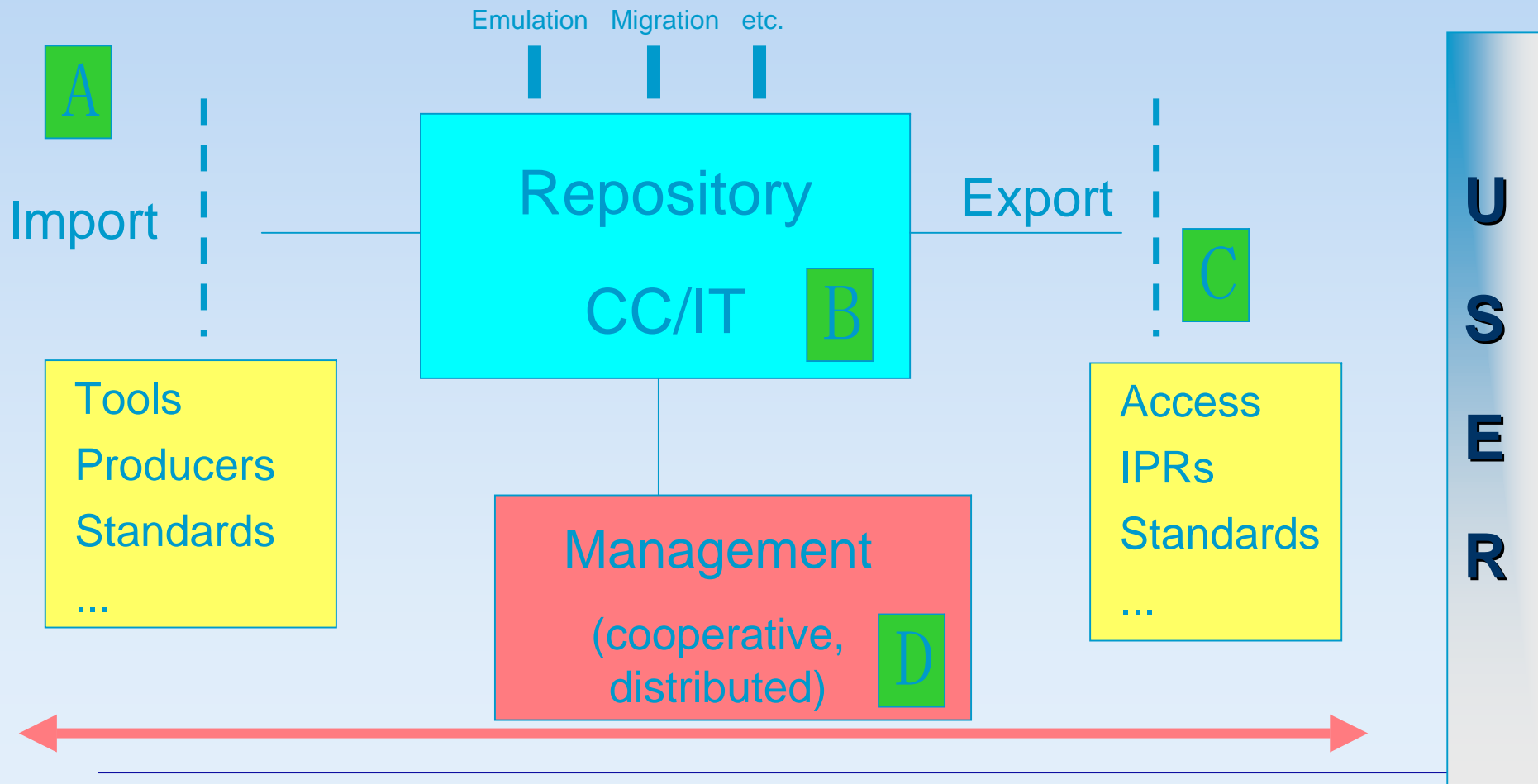
# Model (?)







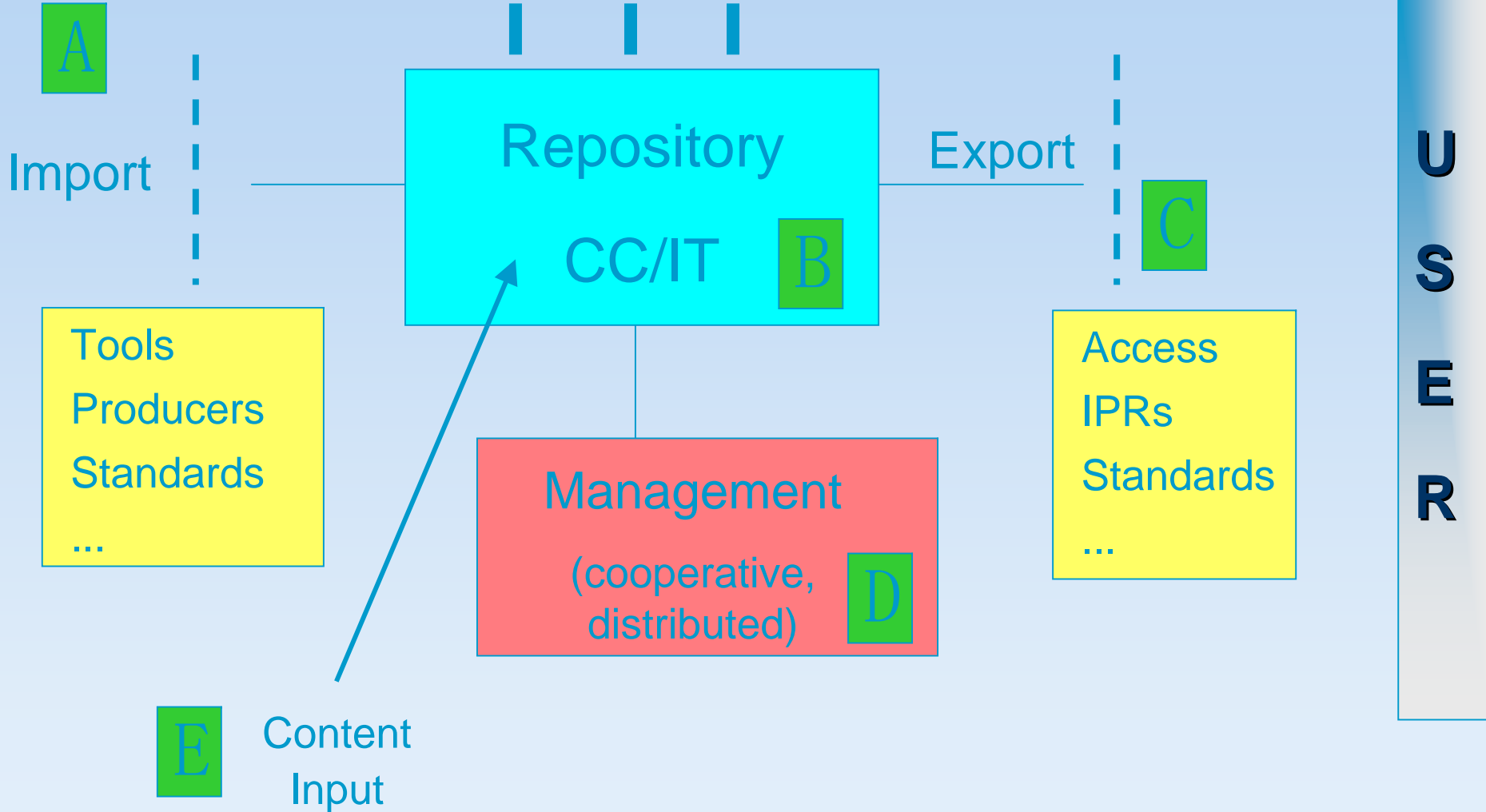
...



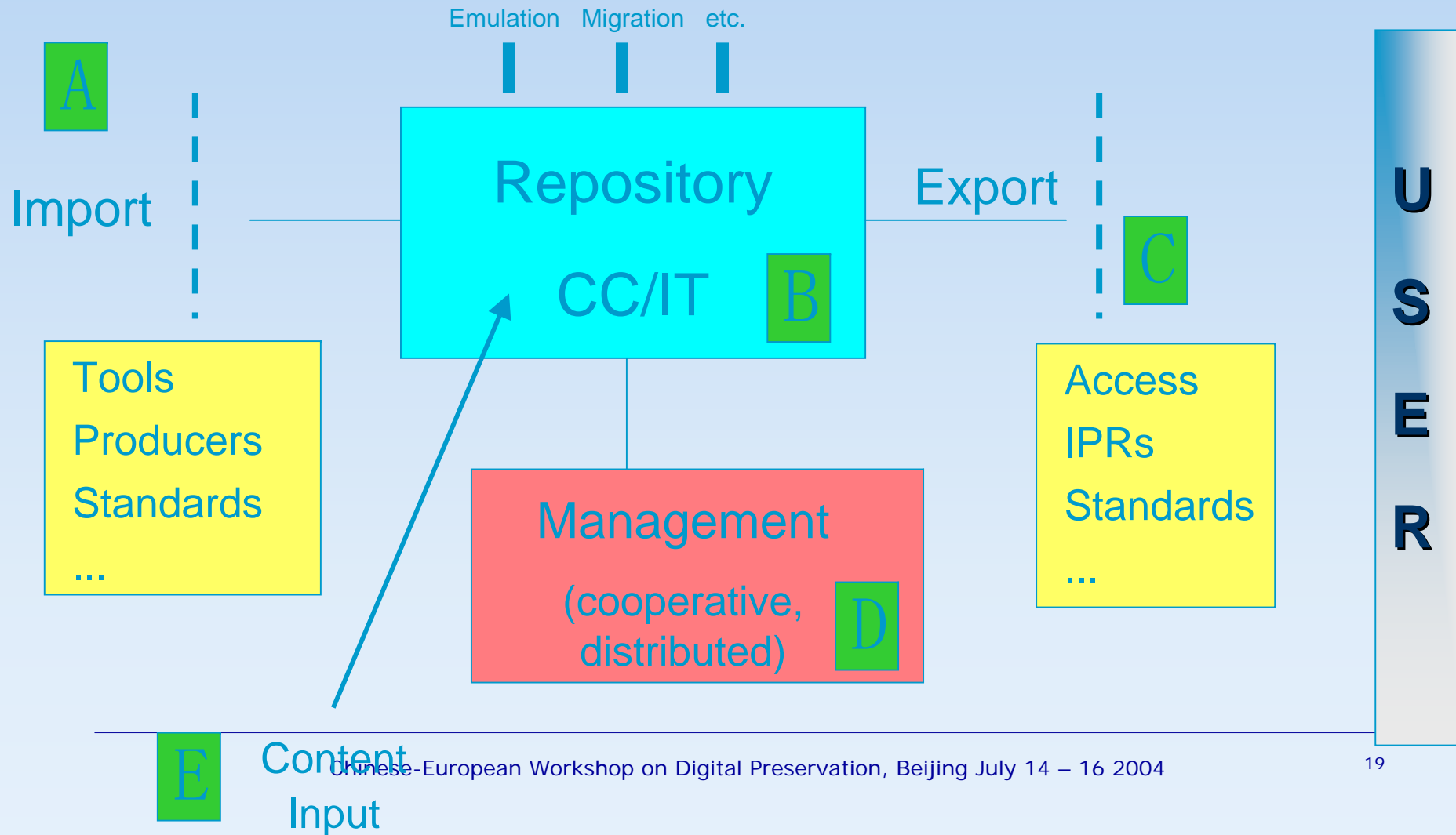


...

Emulation Migration etc.



# PRESERVATION POLICY





## Presentations

- A: Import + E: Content Input
  - Preservation of Image Documents + Case Study  
*René*
  - Preservation of E-Journals and E-Prints  
*Neil*
  - Preservation of Scientific E-Journals: Case Study  
*Hilde*
  - Preservation of Scientific Data in Natural Sciences  
*Thomas*
  - Preservation of Scientific Data in the Humanities  
*René*
  - Preservation of Web Information at the NL China, Case Study  
*Zhigeng*



...

- B: Repository, Archival Storage System
  - Different approaches to Digital Preservation (Migration, Emulation)  
*Hilde*
  - Metadata for Preservation  
*Michael*
  - File Format Characteristics and Significant Properties  
*Andreas A.*
  - The OAIS Reference Model, Current Implementations of the OAIS  
*Michael*
  - Trusted Digital Repositories, Certification  
*Reinhard & Heike*



...

## ➤ C: Export

- Are there already some experiences?

## ➤ D: Management

- Legal Aspects of Digital Preservation  
*Neil*
- File Format Registries  
*Andreas A.*
- Persistent Identifier  
*Reinhard*
- Metadata Registries  
*Heike*



## ➤ Preservation Policies

- Preservation strategies of the NL Netherlands  
*Hilde*
- Using Utility Analysis to Evaluate and Compare Preservation Strategies  
*Andreas R.*
- Preservation Strategy for Third-Party Materials of the Chinese Science Digital Library (CSDL) of the Chinese Academy of Sciences (CAS)  
*Xiaolin*
- Preservation Planning, Institutional Strategies and Policies  
*Thomas*



## Outlook

- Common minimal set of preservation metadata
  - International
  - Standardized
  - Cross-domain
- Distributed management of digital archives, repositories
  - Nobody is able to preserve everything
  - Trusted repositories, certification (registry?)
  - Public preservation policies, strategies (registry?)





...

- Long-term objective: ***Interoperability***
  - Architecture, metadata, exchange format, protocol, ...
- But also:
  - Granularity of digital object?
  - Collection Level Description
  - Digital Rights Management
  - Terminology
  - Sensitisation (e.g. EU)
  - ...
- ***International*** Cooperation, Conferences, Initiatives



国家科技图书文献中心  
National Science and Technology Library

NATIONAL LIBRARY OF CHINA



中国高等教育文献保障系统  
China Academic Library & Information System

CSDL

中国科学院国家科学数字图书馆  
Chinese National Science Digital Library

首页 关于我们 FAQ

Thank you very much  
for your attention



Dr. Heike Neuroth  
Research & Development  
Göttingen State and University Library (SUB)  
neuroth@mail.sub.uni-goettingen.de

Österreichische  
Nationalbibliothek

