e-Infrastructure and digital preservation: challenges and outlook

Reinhard Altenhöner
German National Library, Frankfurt am Main
iPres, 5th October 2009

agenda

- e-Infrastructure: challenging potentials for the DP-community
- Broad-scale strategies and operational solutions
- Key issues: cooperation and integration
- Some conclusions

Strategic level: „e-Infrastructure“ 1

- High level of attention to the topic specifically on research data
- “Infrastructure can be defined as the basic physical and organizational structures needed for the operation of a society or enterprise, or the services and facilities necessary for an economy to function” (Wikipedia)
- Actors (aside of the producer): research institutions, educational institutions (often combined), the (scientific) publication market, facilitators (libraries, professional information providers and data centers)

„e-Infrastructure“ 2

- 7th framework EU-programme (Research infrastructures) ‘ICT Infrastructures for e-Science’: broader understanding of science and research infrastructure:

  “The objective of scientific data e-Infrastructures is to develop an ecosystem of European digital repositories […] to respond to Member State requests to improve access to scientific information. Europe needs to pay particular attention to the accessibility, quality assurance and preservation of key data collections.”

„e-Infrastructure“ 3
German situation with special regard to DP

- 2005 German Research Foundation picked DP into their funding strategy
- 2006: National information infrastructure addresses DP explicitly, but no concrete activities follow
- 2008: Alliance of leading science organizations for „Digital Information“, one topic: ‘national hosting strategy’
- 2009: New national approach to reorganize the national information infrastructure comprising all relevant players, tasks:
  - Identify areas, where additional activities are needed
  - Define and dedicate tasks to institutions
  - Establish nodes of expertise (real / virtual)
  - Recommendations for funding rules, additional measures

„e-Infrastructure“ and DP

- Attention-level is high, potentials are there
- Technology: GRID offers new opportunities
- Funding: Insight that integration of DP-requirements is necessary clearly increases
- But where we are on the solution level?

Solution level w.r.t. DP

- Some solutions are in place (portico, DIAS, kopal, hathiTrust, (C)LOCKSS), some in preparation (SPAR, Rosetta) – mostly institutional, specifically dedicated to DP
- Progress was and is being made especially in EU-funded projects like CASPAR, PLANETS and SHAMAN or in special activities like KEEP

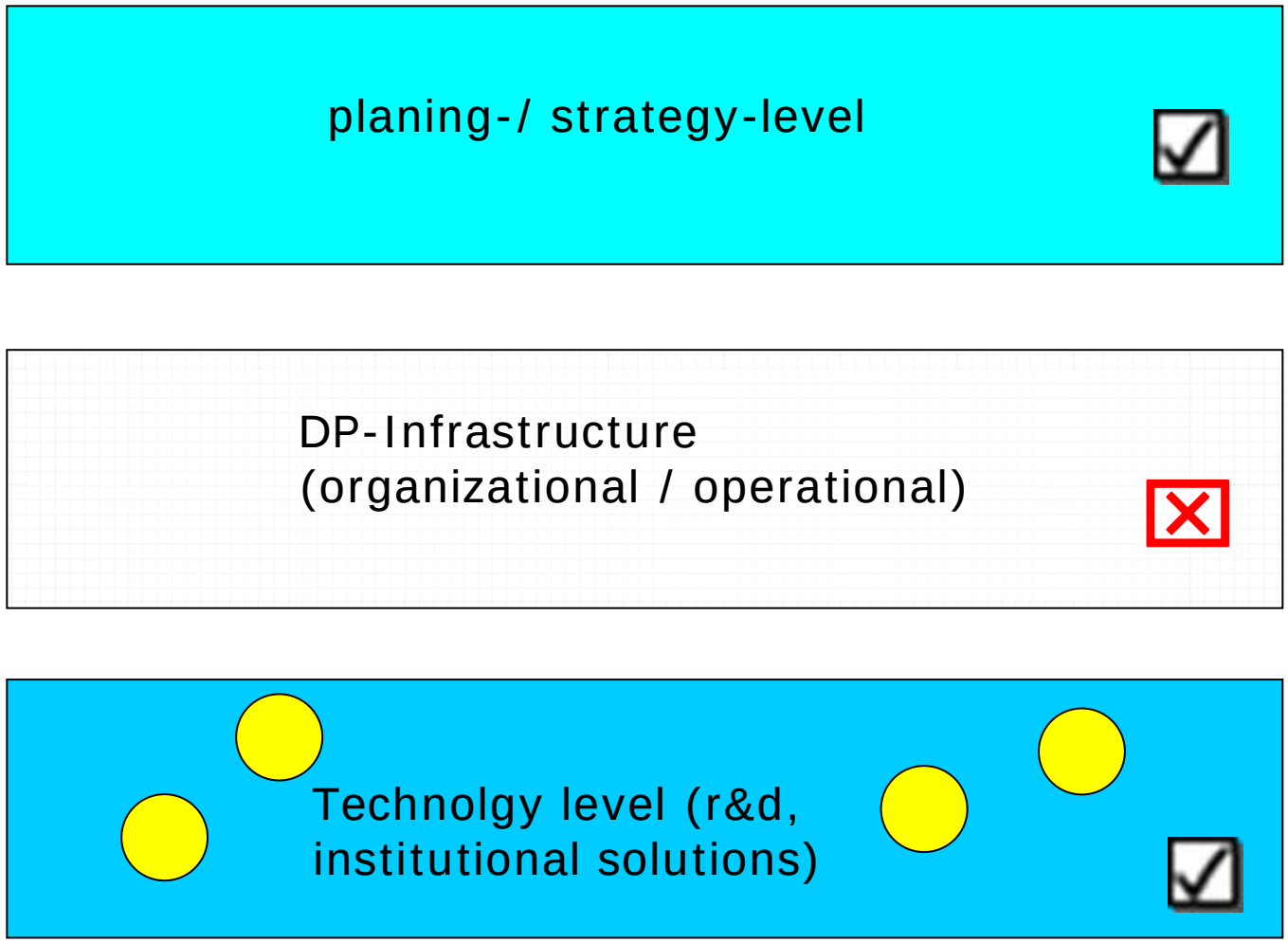
→ There are single initiatives and „safe places“

- But as Digital Preservation Europe (even in 2006) noted: significant mismatch between the scale of the problem and the level of effort being mobilized to address the problem through research

DP-Infrastructure 1

- Parse.Insight (2009) Roadmap: assembles technical and non-technical components aimed at bridging the “islands of functionality, developed for particular purposes [...] separated by discipline or time”
- Significant lack of progress in establishing a common approach to solving the problems of preservation across the spectrum of memory institutions, e.g.
  - different approaches to auditing and certification of trusted repositories
  - many different approaches to preservation-related metadata models
- No clear defined and directed tasks & task sharing, e.g. selection

The path to DP?



DP-Infrastructure 2

- Funding is focused on many individual projects
- We have relevant solutions and approaches, including aspects of infrastructure (PI, format registries), but partly competitive, partly in infancy
  - Who is really doing preservation planning?
  - Who is acting migration / emulation scenarios in a broad scale
  - Where are defined scenarios for risk management and actions?
  - Common Software deposits?
  - Can we confirm that DP is operational?
- No common terminology, no clear idea what DP-services are
- Open questions with regard to the needs (context?!)
- No mechanisms to evaluate DP

DP-Infrastructure 3: other concerns and needs

- Technology, e.g. interoperability, transfer of semantic relations
- Legal matters
- Standardization with regard to formats, workflows, QM
- Collaboration between data producers, indexers, archives and providers
- Collaboration with market-driven players like D21, accenture, broadcasters
- Conversion of different DP-approaches (data archives, memory institutions)
- Professionalization of the DP-community members

Nestor: German network of expertise in long-term storage of digital resources

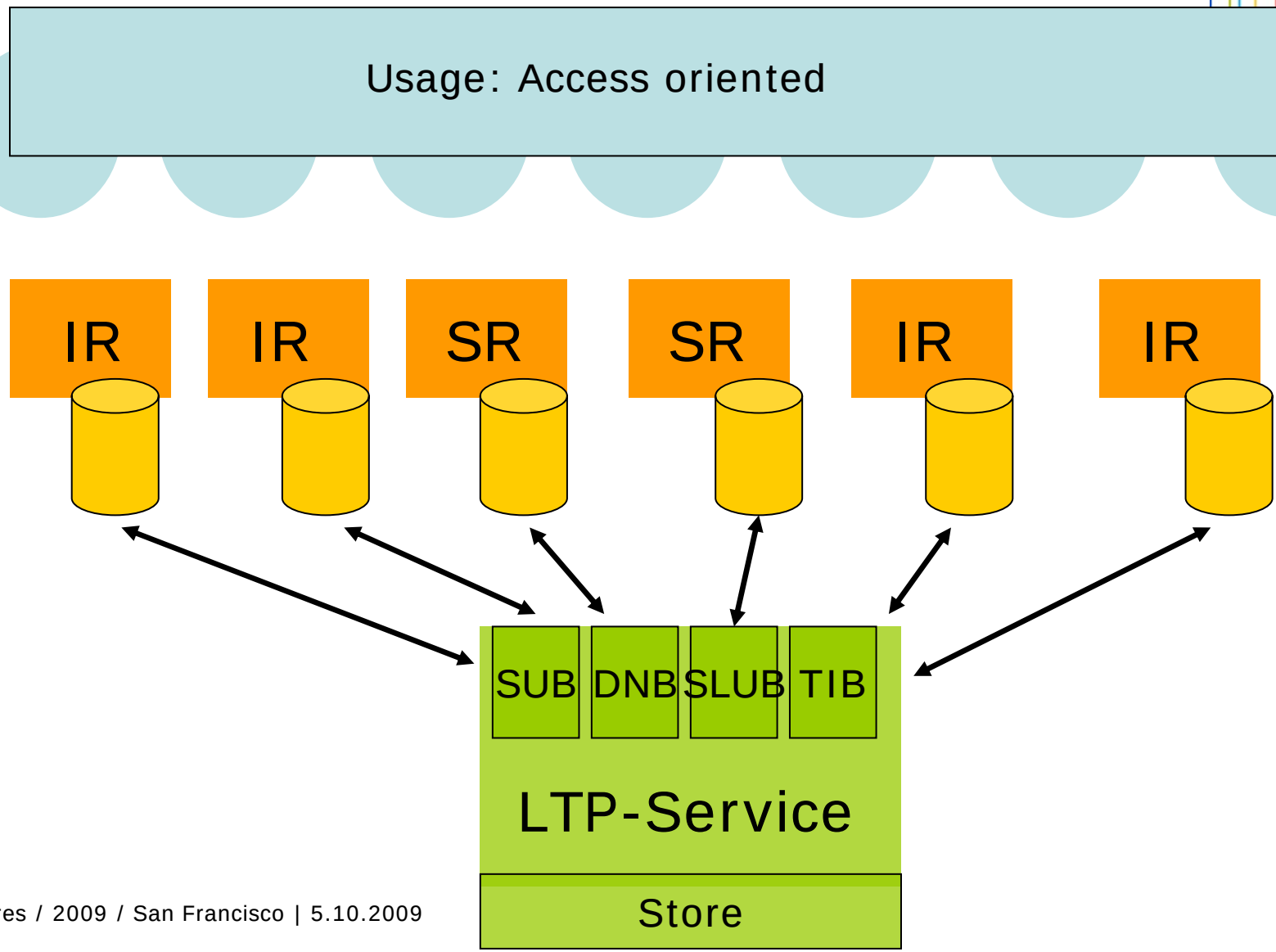
Practical key issues: cooperation and integration

- Time to change: From solutions to well-defined DP-services?
- Technology:
  - Decoupled system-components
  - Separation in services / modular design
  - Open well-defined interfaces
  - OSS / Open licensing situation
  - Transparent documentation

  → Ensuring the capacity to integrate additional functions and services from 3rd parties
- Organization:
  - Institutional cooperation and shared services
  - Integration of DP in the workflow of memory institutions

→ Example

Dp4lib - basic idea



Example: dp4lib 2

DEUTSCHE NATIONAL BIBLIOTHEK

- Cases (in the perspective of the repository) to be adressed:
  - single object is damaged
  - format not accessible
  - Object is missing (policy based)
  - complete collections are damaged or not accessible or missing for some reasons
  - complete restore of the repository on a new platform is needed

Example: dp4lib 1

DEUTSCHE NATIONAL BIBLIOTHEK

- Reuse of existent infrastructure (kopal)
- Characteristics kopal:
  - Core (DIAS): extended with remote access functionality, allows independent use of one system for different customers
  - Bit stream preservation is outsourced
  - Well defined SIP / DIP interface (Universal Object Format)
  - Locale software: Flexible Java modules to generate and extract different metadata, to create the universal object format and to integrate ingest and access in existing software environments
    → Open source
  - Toolbox to generate technical metadata

Example: dp4lib 3

- Todos:
  - Extension of technical interfaces (selection procedures, dedicated access control, transferring object information)
  - Specific services of the DP-store must become defined and implemented.
  - Extension of metadata-extraction and format-evaluation tools, esp. in the area of multimedia (movies), clickable's
  - Definition of segmented generic work processes to implement digital preservation on the repository-level
  - „Overhead" – mechanism to govern the definition of customized service models for DP

Conclusions

- The broad-scale discussion on e-Infrastructure shows: the needed level of attention is obtained
- There is a gap between the approach for a DP infrastructure and the available services
  - We need national/international corporate bodies for DP (facilitate national / international coordination and corporation, task mapping, funding)
  - We need concrete steps for a global infrastructure for registries, data formats, Software deposits, risk management
- We need to improve practice examples of work share (technology, organizational level)
- We need more effort on technical and organizational workflow integration

Reinhard Altenhöner

mailto:r.altenhoener@d-nb.de
http://www.d-nb.de