AVOIN TIEDE JA TUTKIMUS

Preservation of scientific information in open science era

PhD Pirjo-leena Forsström, Development Director
Secretary-General of Finnish Open Science and Research Initiative

INDEX

- On digital preservation
- Managing digital preservation
- Open development
- Open science

Definitions

- Curation : The activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose.
- Archiving : A curation activity which ensures that data is properly selected, stored, can be accessed and that its logical and physical integrity is maintained over time, including security and authenticity.
- Digital Preservation : An activity in which digital objects and information are maintained over time so that they can still be accessed and understood through changes in technology.
- Digital curation : looking after and somehow "adding value" to digital data, ensuring its current and future usefulness. This probably implies creating some new data from the existing, in order to make the latter more useful and "fit for purpose".

Digital preservation of research results

AVOIN TIEDE JA TUTKIMUS

- Preservation of digital information is at the core of research process => concerns all research organizations
- There is a growing need for digital preservation of research information for several decades or even hundreds of years.
- Today, there is been no controlled and guaranteed way of handling digital information in the long term.
- Equipment, software, and file formats will become outdated, but despite this the information must be preserved in an understandable form (validation and verification needs, re-use)

Preservation methods

AVOIN TIEDE JA TUTKIMUS

Preserving the original look-and-feel

- Emulation
  - Development of emulators to new platforms etc.
  - Active testing and technology watch

Preserving the content

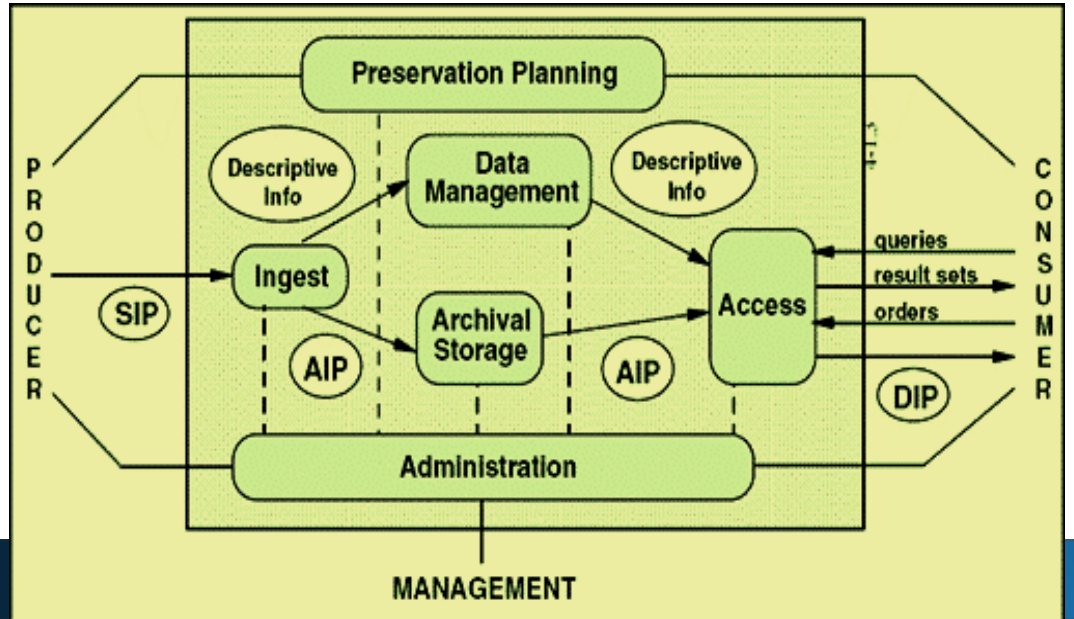
- Migration
  - Format development watch (format libraries)
  - Development of transformation processes, testing, implementation, monitoring
  - Preparation for recoveries

Preserving the bits

- Integrity
  - File validation and monitoring
  - Management of copies
  - Both objects and metadata

Digital Preservation Framework

- the ISO OAIS Reference Model for *an* OAIS. This reference model is defined by recommendation CCSDS 650.0-B-1 of the Consultative Committee for Space Data Systems;[1] this text is identical to ISO 14721:2003.



Source: Long-Term Preservation of Digital Documents. 2006. doi:10.1007/978-3-540-33640-2. ISBN 978-3-540-33639-6. Public Domain.

What should be preserved?

AVOIN TIEDE JA TUTKIMUS

- For validation, all relevant information from the research process:
  - Publications
  - Data
  - Methods
  - Metadata, quality information
  - References, linkages
  - IPR and ownership information, license

Risk management

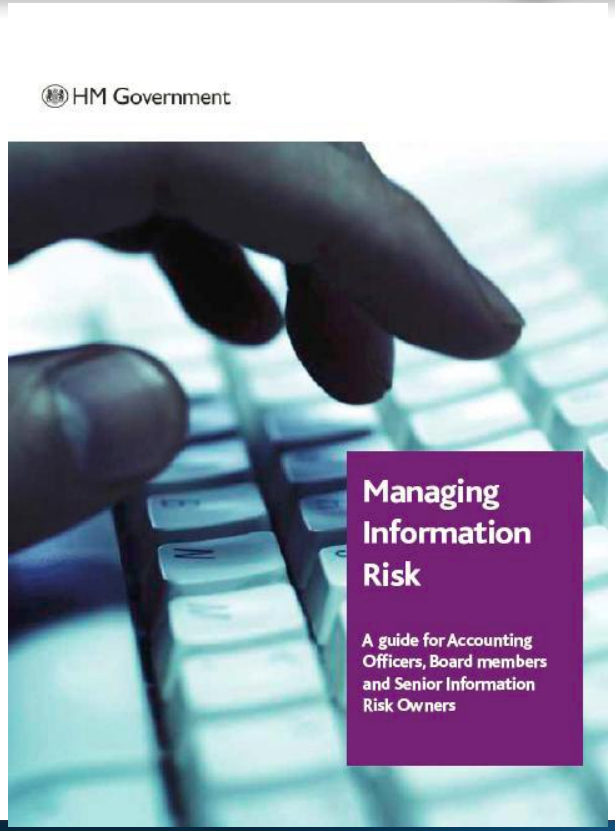
AVOIN TIEDE JA TUTKIMUS

- Media failure and obsolescence
- Hardware failure and obsolescence
- Software failure and obsolescence
- Communication errors
- Failure of network services
- Operator error
- Natural disaster
- External Attack
- Economic and organizational failure

(Rosenthal et al., 2005)

Information risks in digital preservation

AVOIN TIEDE JA TUTKIMUS



Risk category	Example of risk
Governance and culture	Lack of comprehensive oversight and control (so anything can go wrong) When something goes wrong, handling it badly and not learning (so it can happen again) Third parties let you down (letting down your customers and your reputation suffers) New business processes don't take information risk into account (with serious consequences)
Information management and information integrity	Critical information is wrongly destroyed, not kept or can't be found when needed (leading to reputational damage or large costs) Lack of basic records management disciplines (can have wide-ranging consequences) Inaccurate information (which causes the wrong decision to be made, or the wrong action to be taken) Vital electronic information becomes unreadable due to technical obsolescence (with legal, reputational or financial consequences) Critical information is lost (with legal, reputational or financial consequences)
The human dimension	Despite having procedures and rules, staff, acting in error, do the wrong thing (and things go badly wrong) Despite having procedures and rules, 'insiders', acting deliberately, do the wrong thing (and things go badly wrong) External parties get your information illegally (and expose it/act maliciously/defraud you or your customers)
Information availability and use	Inappropriate disclosure of sensitive personal information (causing reputational damage or worse) Failure to disclose critical information for case management/protection (at worst leading to loss of life) Failure to utilise the value of the information asset (leading to a waste of public money) Failure to allow information to get to the right people at the right times (leading your service to fail your customers)

Demands for preservation architecture

AVOIN TIEDE JA TUTKIMUS

- Digital objects are copied to different medias
- Data integrity watch is a constant process
- The system scales up to Big data volumes
- Critical process steps are duplicated (at least)
- Preservation process is geographically distributed
- Architecture is based on components (both Hw and SW)
- Hardware and software layers are constantly monitored, and new componets are added and old removed according to need
- Governance is systematic, well organized, anticipatory, person-indipendent, transparent and traceable

Governance in digital preservation

AVOIN TIEDE JA TUTKIMUS

- Preserving scientific information: at the core of research process
- Involves science, technology, and innovation issues
- Addressing such complex issues calls for effective governance mechanisms
- There are no simple solutions
  - Good governance practices

Governance in digital preservation: lessons learned

AVOIN TIEDE JA TUTKIMUS

- Institutional framework for priority setting should be flexible
- Flexible funding and spending mechanisms help ensure stability
- Knowledge sharing and intellectual property require tailored approach
- Outreach is indispensable for putting preservation into practise

Checklist for policy options

AVOIN TIEDE JA TUTKIMUS

- *The importance of high-level co-ordination of the project.*
- *Need for a compelling reason to do the work.* Where a link to high-level political commitments cannot be made, a demand-led approach seems most promising. Co-operation should focus on fields with clear knowledge needs shared by many decentralised actors who perceive clear benefits to international co-operation when compared to acting on their own.
- *The governance structure must be a "learning system".*
- *System linkages are important.* Linkages should seek to include a broad and relevant range of stakeholders while maintaining an effective decision-making process.
- *Outreach and knowledge flows outside the project.*

Cheklist for policy options continued

AVOIN TIEDE JA TUTKIMUS

- *Knowledge flows and knowledge protection.* Knowledge sharing and IP provisions should be adapted as necessary to each phase of the collaboration life cycle. This is particularly important given that IP issues tend to increase in importance as a product nears market deployment.
- *Contingency management.* Funding and spending mechanisms should contain contingency provisions. In the case of delayed payments, or the need to fund multi-annual projects with annual funding, mechanisms are needed to provide for funding and spending stability.
- *Combining co-operation with capacity building.* Capacity building is an important element of joint efforts to address these challenges, and should not be seen as a support mechanism only, or not even mainly,

Breaking points in governance

AVOIN TIEDE JA TUTKIMUS

- Primary documentation
- Commitments
- Capabilities
- Resources
- Strategies

Research process

Open development in digital preservation

AVOIN TIEDE JA TUTKIMUS

- Good ideas are widely distributed today, no one has a monopoly on useful knowledge
- Innovation is now done within networks, rather than within a single firm
- Not all of the smart people in the world work for us

Open preservation framework:

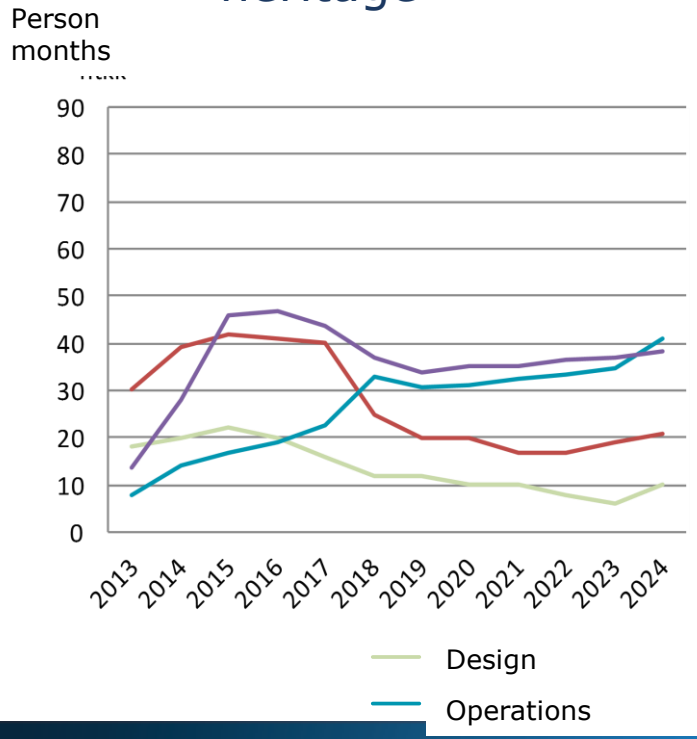
- Ability to profit from technology
- Ability to scale up technology
- Ability to continue innovating technology
- Ability to acquire technology
- Ability to involve new skills

Open Planets Foundation
http://www.openplanetsfoundation.org/

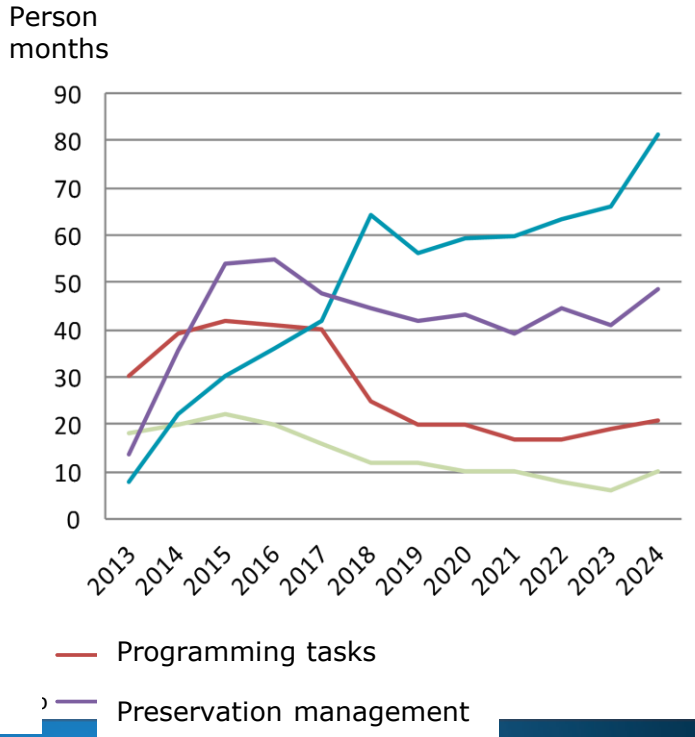
Workload distribution

AVOIN TIEDE JA TUTKIMUS

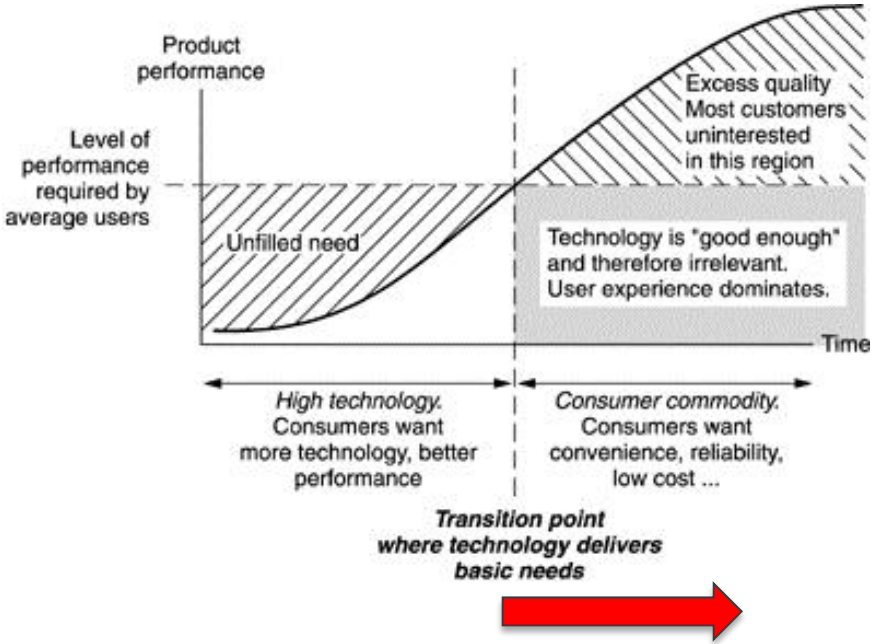
Preservation of cultural heritage



Preservation of scientific data



Open development in digital preservation

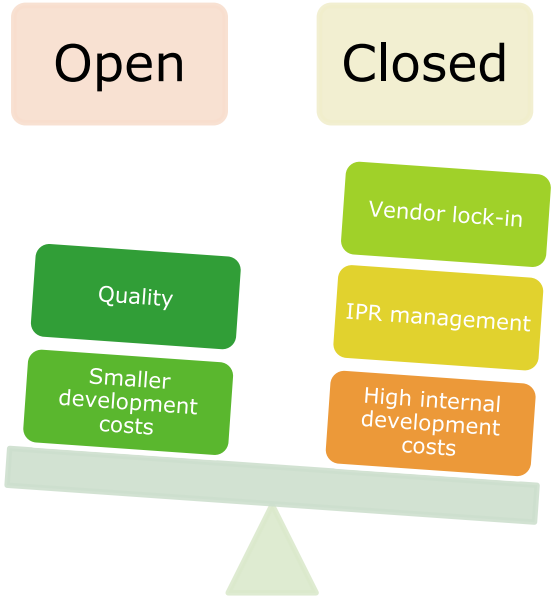


Development of next generation building block

From: Christensen, C. M. (1997). The innovator's dilemma: When new technologies cause great firms to fail. Boston: Harvard Business School Press.

Open code in digital preservation

- Open code facilitates continuity planning of software componets.
- Open and standard file formats are important in preserving content.
- Most important: open standards for connections and communication between data and software, extentensive use of standards
- rising costs of technology development, if not done openly with others
- shortening idea life cycle, when struggling with infrastructure takes valuable research time

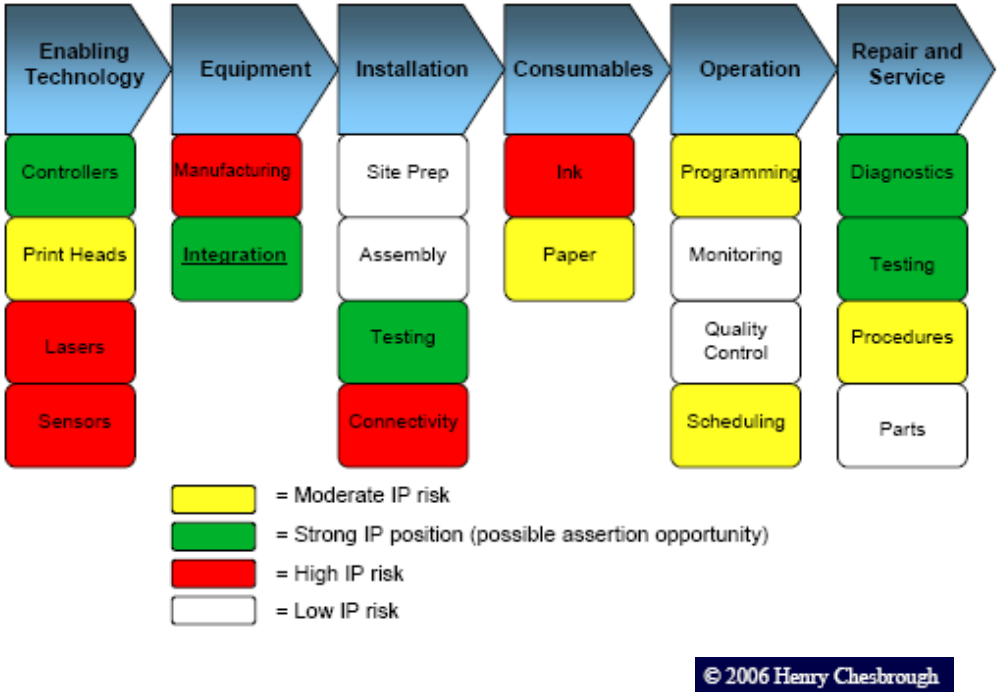


Open and closed code in digital preservation

AVOIN TIEDE JA TUTKIMUS

IP risks

- Example: printers



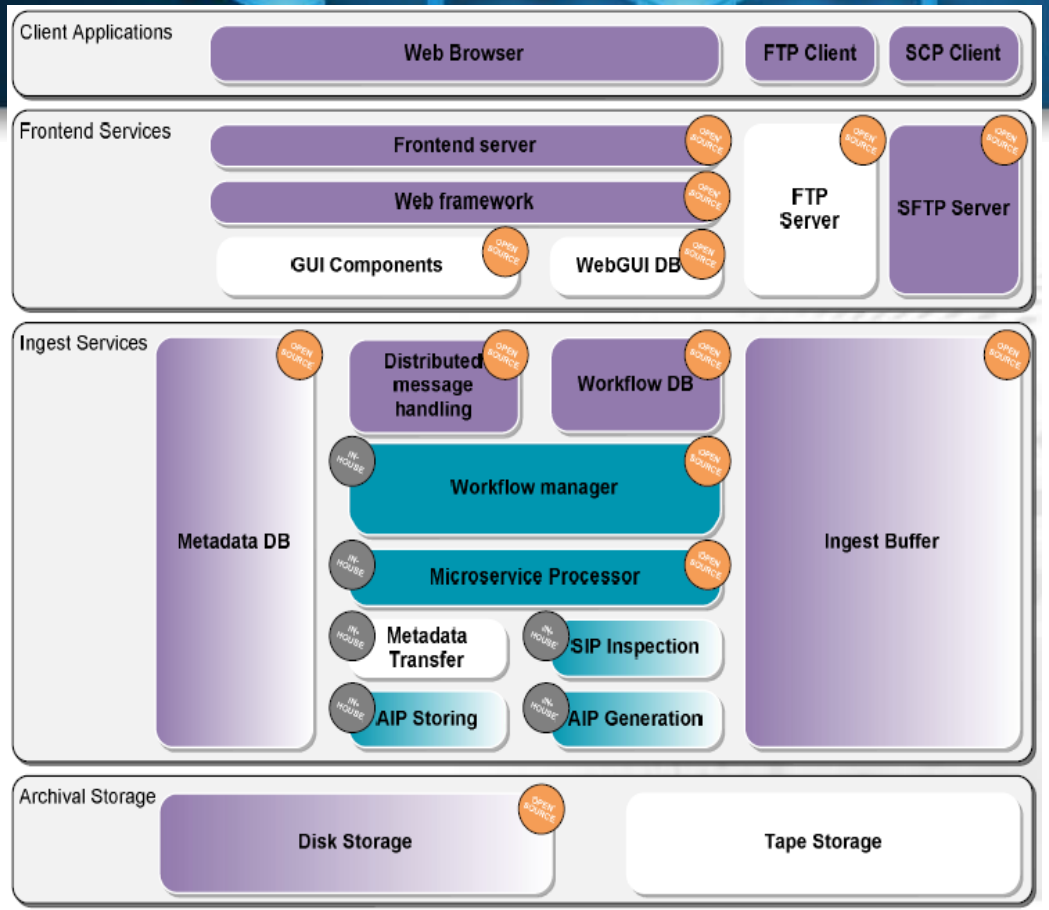
Example: Ingest

AVOIN TIEDE JA TUTKIMUS

- We utilize 24 different Open Source components
  - Format checks: 11 components (JHOVE1, JHOVE2, FITS, Epubcheck, Apache ODF Toolkit, Officetron, FLAC, Pngcheck, warc-tools, Ms Office binary File Format Validator, MP3val)
- Missing parts done in-house
- Lots of integration work (technology watch, testing, report handling etc.)

Open strategy:

Components licenced with open licenses


Client Applications
Web Browser
FTP Client
SCP Client
Frontend Services
Frontend server
Web framework
GUI Components
WebGUI DB
FTP Server
SFTP Server
Ingest Services
Metadata DB
Distributed message handling
Workflow DB
Workflow manager
Microservice Processor
Metadata Transfer
SIP Inspection
AIP Storing
AIP Generation
Ingest Buffer
Archival Storage
Disk Storage
Tape Storage

On the openness of science and research

AVOIN TIEDE JA TUTKIMUS

- Open science is an **internationally** significant way of promoting science and the **impact** of science in society.
- Openness is a core principle of science and research, through which new possibilities for **engaging** in science and research are created for scientists, decision-makers and the citizens.
- It requires the wide-scale **availability** of the publications, data, methods, know-how and support services produced and required by research.
- Digitalising and opening research processes creates **new opportunities** for cooperation and communication for researchers and stakeholdes. Increases credibility of science and promotion of innovation.

Policy proposal concerning open access of research results in Finland

AVOIN TIEDE JA TUTKIMUS

- Basic objective
  - Research data and publications are openly available in an information network via an open interface
- Clarifications (extracts)
  - All actors in the Finnish research system share the scientific publications and research data they produce through an open information network. This principle of openness also concerns research methods and the tools required to produce results, such as computer simulations.
  - Openness will, however, adhere to ethical principles and respect the legal context. Open access to research data will always be the goal when it is legally and contractually possible.
  - The re-use of research data and publications is not unnecessarily restricted, and the terms and conditions of their use are clearly stated. Standard, generic and machine-readable licences are complied with - for example, CC BY 4.0, which will be receiving a Public Administration Recommendation (JHS).
  - The contracts and funding decisions that concern research, support open access to publications and data.

Open science

AVOIN TIEDE JA TUTKIMUS

The new market for scientific information has to work

- outside-in: bringing ideas effective to research process
- inside-out: enable others to utilize unused ideas

Open science is by nature:

- open-ended
- includes multi-stakeholders
- transformative

Open science enablers

AVOIN TIEDE JA TUTKIMUS

- Clear guidelines concerning **ethical principles, IPRs and legal constraints**. Legislative environment should encourage open science, and no uncertainty should linger if results should be open or not. The ethical principles for research should be clearly stated and easily consulted. Legal and IPR issues should be mandatory to resolve prior to research.
- **Licensing** of research results via open license (like Creative commons 4.0 BY) should be demanded.
- **Service infrastructure** for open science should be clearly defined, and using it should be mandatory. This involves common services like researcher identification, persistent identifiers, preservation and archival services. Service infrastructure components have to be **interoperable**. **Open source code, open standards and open interfaces** should be used in the infrastructure.

Conclusions

- If you want your digital data to survive, start today!
- Risk management:
  - Governance and policies
  - Open development
  - Look at the whole open science process
- Collaboration
- Pro-active design

AVOIN TIEDE
JA TUTKIMUS

Thank you!

More info: openscience.fi

Pirjo-leena.forsstrom@csc.fi