

The what, why and how of long-term data preservation



Ingrid Dillo

Deputy Director DANS/Member of RDA TAB

e-IRG Workshop Long-term Sustainability

Malta, 9 June 2017





DANS organisation

Mission: promote
and provide
permanent
access to digital
research
resources

Institute of
Dutch Academy
and Research
Funding
Organisation
(KNAW & NWO)
since 2005

First predecessor
dates back to
1964 (Steinmetz
Foundation),
Historical Data
Archive 1989

DANS core services

<https://dans.knaw.nl>



EASY

Certified Long-term Archive



DataverseNL
to support data
storage during
research until
10 years after

NARCIS

Portal
aggregating
research
information and
institutional
repositories

DANS international connections



CESSDA

Council of European Social Science Data Archives. »»



DARIAH

Digital Research Infrastructure for the Arts and Humanities. »»



DCCD

Digital Collaboratory for Cultural Dendrochronology »»



EHRI

European Holocaust Research Infrastructure. »»



EOSC pilot

Pilot European Open Science Cloud »»



EUDAT

European data Infrastructure for scientific research »»



HaS-DARIAH

Humanities at Scale »»



K-PLEX

Knowledge Complexity »»



KNOWeSCAPE

European collaboration in visualising knowledge. »»



OpenAIRE2020

Open Access Infrastructure for Research in Europe. »»



PAN

Portable Antiquities of the Netherlands »»



Parthenos

Pooling Activities, Resources & Tools for Heritage E-research. »»

DANS international connections



Proliferation of data

- Growing recognition of the value of data
- Trend of open science/open data/data sharing
- Funders mandate data stewardship

Advantages

- Transparency and replication of research (scientific integrity)
- Reuse of data (efficiency, return on investment, standing on the shoulders of others)



Events

Apr Open Science Conference

04 Expert or political meeting | Ministry of Education, Culture and Science | Competitiveness

4 April 2016 - 5 April 2016
Europe Building, Amsterdam

Open Science is a key priority of the Dutch Presidency. The Netherlands is committed to open access to scientific publications and the best possible re-use of research data, and it would like to accelerate the transition this requires.



INTERNATIONAL ACCORD ON OPEN DATA FOR OPEN SCIENCE

Preface

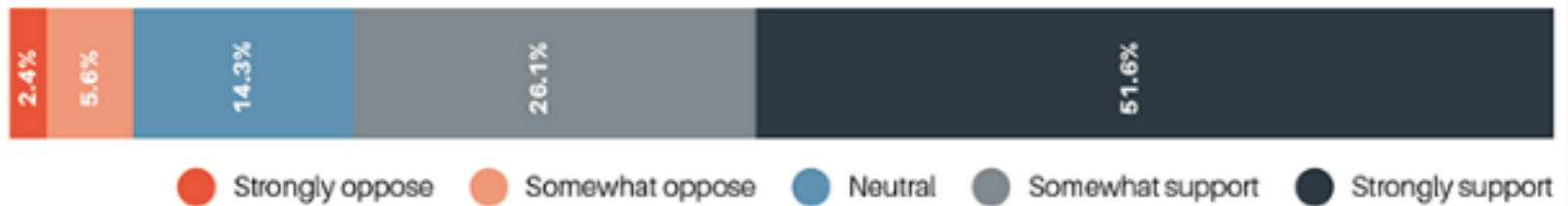
*The International Council for Science (ICSU), the Inter-Academy Partnership (IAP), The World Academy of Sciences (TWAS) and the International Social Science Council (ISSC) have created a joint enterprise, **Science International**, to be the global science community's voice of policy for science. This accord is its first foray in that domain. The accord identifies the challenges and opportunities of the global data revolution as the predominant issue of current policy for science. It wishes to add the distinctive voice of the scientific community to those of governments and inter-governmental bodies that have made the case for open data as a fundamental pre-requisite if science is to maximise its public benefit from the data revolution. It builds on ICSU's 2014 statement on open access by endorsing the need for an international framework of principles of open data as set out in the following document.*

Policy makers



..but what about the researchers?

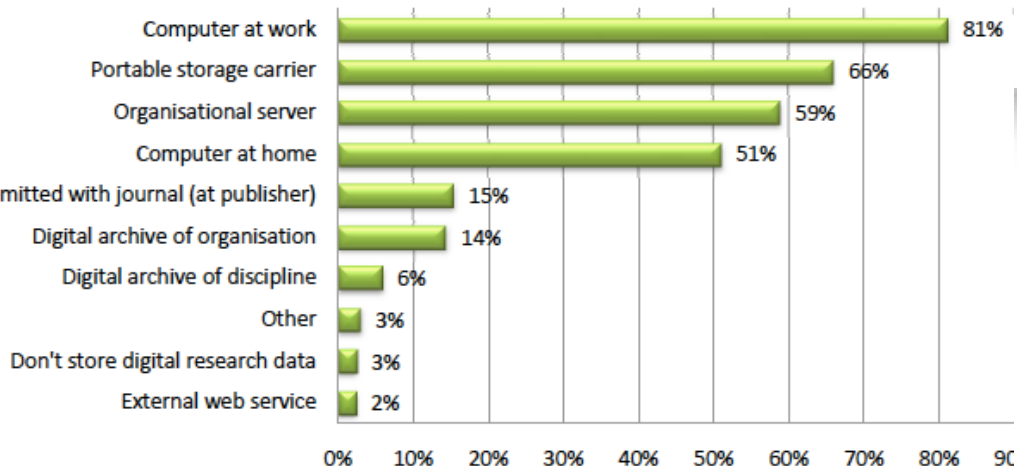
How supportive would you be of a national mandate for open data?



Source: [The State of Open Data, Digital Science Report](#) (2016). Retrieved: December 23, 2016 . Figures have been redrawn from the originals.

Where do you as a researcher store your data for future use?

Multiple answers allowed



Hesitance in reality



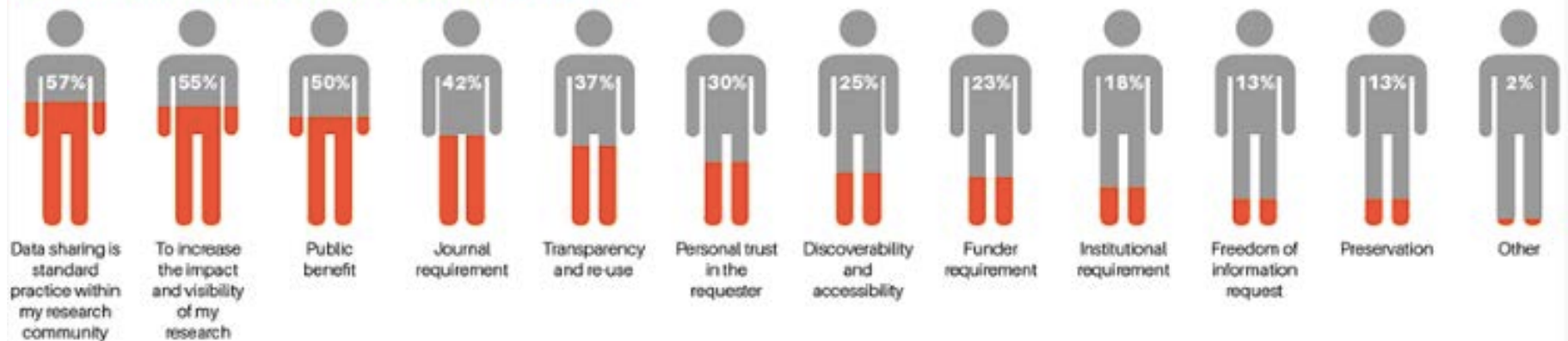
SCENE FROM THE PAST?

Reasons why researchers are hesitant to share their data



Motivations for data sharing

Researcher motivations for sharing data



Source: Wiley's Research Data Insights Survey (2014).

Retrieved: December 23, 2016 . Figures have been redrawn from the originals.

Data sharing incentives

- Influence of sharing norms within direct research circle
- Professional rewards for data sharing
- External drivers:
 - Publisher requirements (DAPs)
 - Funder policies/mandates



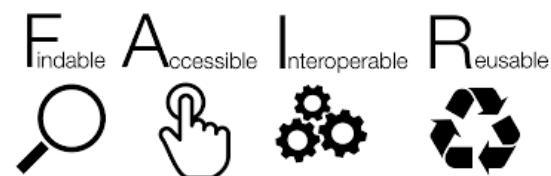
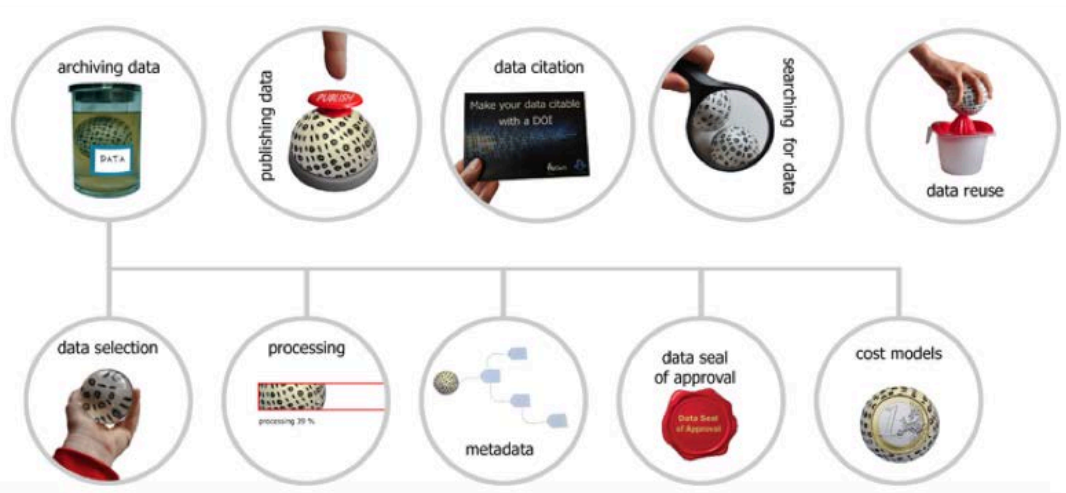
http://repository.jisc.ac.uk/5662/1/KE_report-incentives-for-sharing-researchdata.pdf



Other data sharing challenges

Enabling the researcher to comply with open data requirements:

- awareness raising, training and support for data management (DMPs, FAIR data)
- infrastructure for preservation of and long-term access to the data



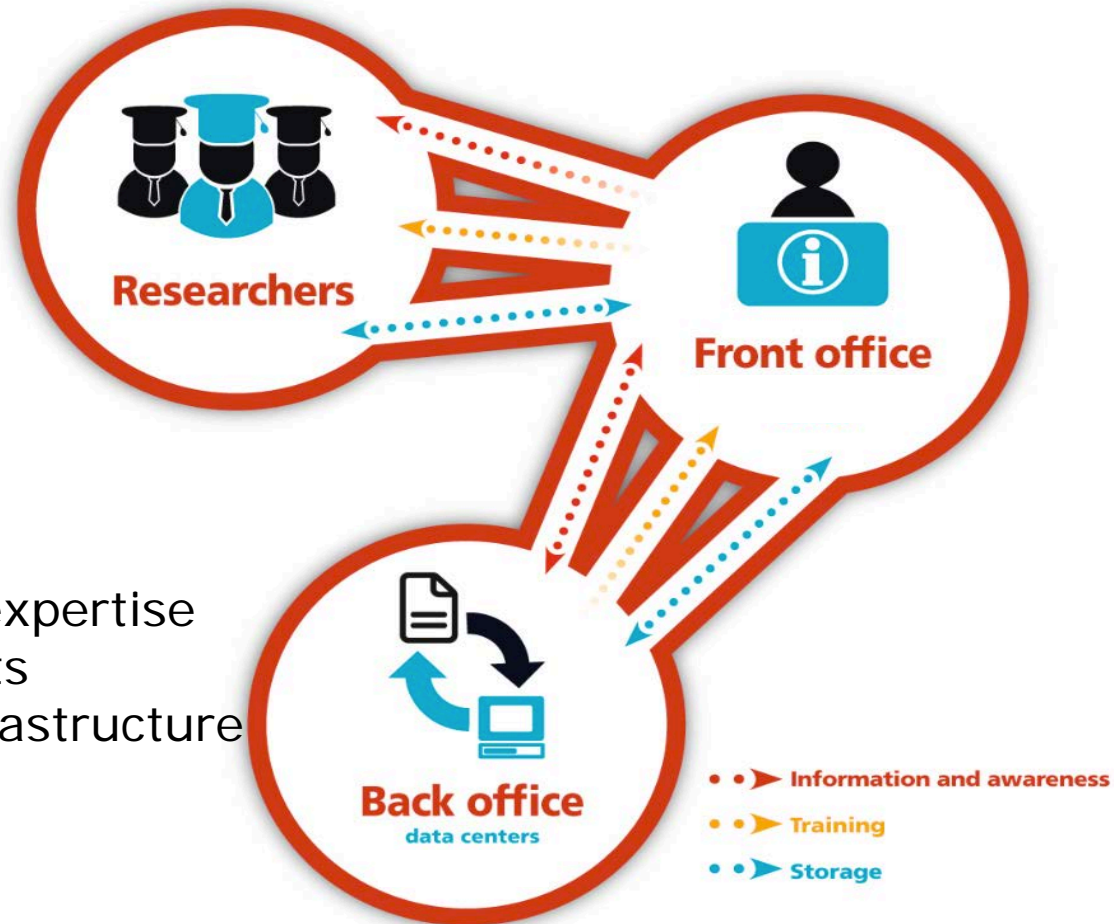
Sustainable support model


Frontoffice-backoffice model

- Division of labour
- Economies of scale

Backoffice

- Curation and preservation expertise
- Training of local data experts
- Long-term preservation infrastructure





“Perhaps the biggest challenge in sharing data is trust: how do you create a system robust enough for scientists to trust that, if they share, their data won’t be lost, garbled, stolen or misused?”

The Data Harvest:

How sharing research data can yield knowledge, jobs and growth

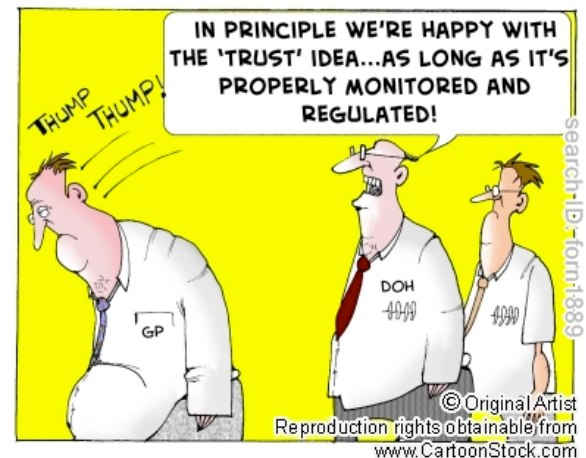
An RDA Europe Report

December 2014

Pillars of trust

- actions and attributes of the trustee (integrity, transparency, competence, predictability, guarantees, positive intentions)
- external acknowledgements:
 - reputation (researchers)
 - third party endorsements (funders, publishers)

What are They
Saying
About
You?



The global certification landscape



DANS and Data Seal of Approval



- 2005: DANS to promote and provide permanent access to digital research resources
- Formulate quality guidelines for digital repositories including DANS
- 2009: international DSA Board
- Almost 70 seals acquired around the globe, but with a focus on Europe
- <https://www.datasealofapproval.org/en/>

Partnership with WDS under the umbrella of RDA

- Goals:
 - Realizing efficiencies
 - Simplifying assessment options
 - Stimulating more certifications
- Outcomes:
 - Common catalogue of requirements for core repository assessment
 - Common procedures for self-assessment and review process
 - One new certification body: **CoreTrustSeal** Board



New CoreTrustSeal Requirements

Requirements:

- Context (1)
- Organizational infrastructure (6)
- Digital object management (8)
- Technology (2)




- LIFE 1, 2 and 3. Projects to explore digital preservation costing, and develop costing models.
 - <http://www.life.ac.uk/>
- Cost Model for Digital Preservation (CMDP): Project at the Royal Danish Library and the Danish National Archives to develop a new cost model. Currently covers Planning, Migrations and Ingest
 - <http://www.costmodelfordigitalpreservation.dk>
- Keeping Research Data Safe 1 and 2 (KRDS): Cost model and benefits analysis for preserving research data
 - <http://www.beagrie.com/krds.php>
- Presto Prime cost model for digital storage
 - <http://prestoprime.it-innovation.soton.ac.uk/>
- Cost Estimation Toolkit (CET): Data centre costing model and toolkit, from NASA Goddard
 - http://www.pv2007.dlr.de/Papers/Fontaine_CostModelObservations.pdf
- Cost Model for Small Scale Automated Digital Preservation Archives (Strodl and Rauber)
 - http://www.ifs.tuwien.ac.at/~strodl/paper/strodl_pres2011_costmodel.pdf
- APARSEN Project activity focused on digital preservation costing
 - <http://www.alliancepermanentaccess.org/index.php/knowledge-base/digital-preservation-business-models/costbenefit-data-collection-and-modelling>
- EPRSC and JISC study on Cost analysis of cloud computing for research
 - http://www.jisc.ac.uk/media/documents/programmes/research_infrastructure/costcloudresearch.pdf
- Cost forecasting model for new digitization projects (Excel and web tool under development) (Karim Boughida, Martha Whittaker, Linda Colet, Dan Chudnov)
 - http://www.cni.org/wp-content/uploads/2011/12/cni_cost_boughida.pdf
- DP4lib is developing a business and cost model for a digital preservation service
 - <http://dp4lib.langzeitarchivierung.de/downloads/DP4lib-One-Pager-08-eng.pdf>
- DANS Costs of Digital Archiving Volume 2 Project, focusing on preservation and dissemination of research datasets
 - <http://www.dans.knaw.nl/en/content/categorieen/projecten/costs-digital-archiving-vol-2>
- Blue Ribbon Task Force on Sustainable Digital Preservation and Access
 - <http://brtf.sdsc.edu/>
- Economic Sustainability Reference Model (draft)
 - <http://4cproject.eu/community-resources/outputs-and-deliverables/ms9-draft-economic-sustainability-reference-model>
- ENSURE Project - Enabling kNowledge Sustainability Usability and Recovery for Economic value
 - <http://ensure-fp7-plone.fe.up.pt/site>
- 4C. EU funded project on costing, led by JISC.
 - <http://4cproject.eu/>
- Cost Model for Electronic Health Records (Bote, Fernandez-Feljo, and Ruizb)
 - <http://www.sciencedirect.com/science/article/pii/S2212017312004434>
- TCP: Total Cost of Preservation (California Digital Library)
 - <https://wiki.ucop.edu/display/Curation/Cost+Modeling>
- Cost model for digital preservation (National Archives of the Netherlands)
 - http://dlmforum.typepad.com/Paper_RemcoVerdegem_and_JS_CostModelfordigitalpreservation.pdf



The cost of long term preservation





Curation Costs Exchange

Understanding and comparing digital curation costs to support smarter investments


[Home](#) [About](#) [Understand costs](#) [Compare costs](#) [Read more](#) [Vendor services](#) [Help](#) [Sign Up](#) [Sign In](#)


All about the costs of curation

What am I spending, what are they spending, what should we be spending?

Understand costs

Assessing your costs and using cost models to make smart investments





Compare costs

Add your curation costs and see how they compare with others

Sustainable business models for data repositories

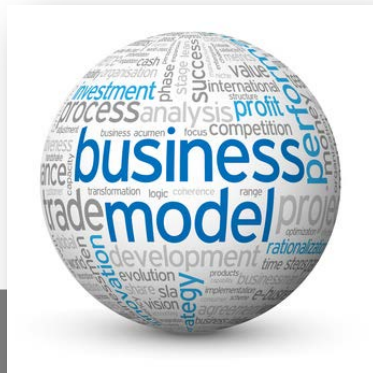
Increasing need for data repositories and data stewardship.

- Increasing volume presents a challenge.
- Requirements for stewardship present a greater challenge.



Sustaining digital data infrastructure is a major issue for science policy

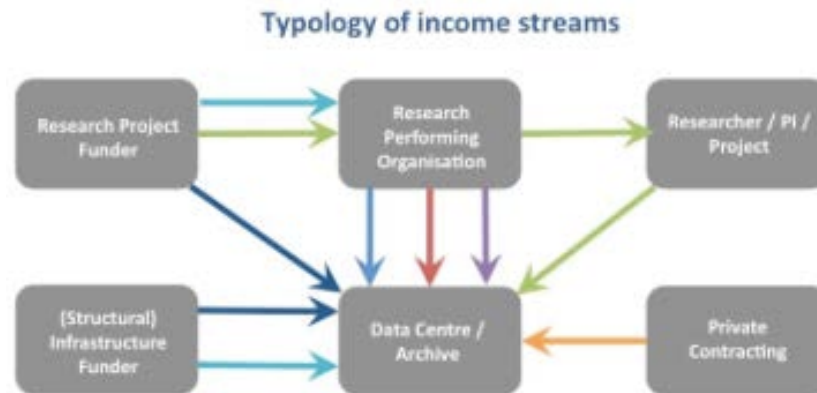
- current funding models will prove inelastic and not meet the growing requirements – concern on the part of repositories and funders



Sustainable business models for data repositories

RDA Cost Recovery Interest Group, also supported by WDS and CODATA
Report *Income Streams for Data Repositories* (Feb 2016;
<https://zenodo.org/record/46693#.WTUR-TOB2T8>)

- based on 25 in-depth interviews, identifying topics and trends, alternative revenue streams



- 1) Structural (central contract)
- 2) Hosting Support (indirect or direct support through institutional hosting)
- 3) Annual Contract (from depositing institution)
- 4) Data Deposit Fee (may be paid by researcher, RPO or publisher; may originate with funder)
- 5) Access Charge (for the data or for value-adding services)
- 6) R&D Projects (to develop infrastructure or value-adding services)
- 7) Private Contracting (services to parties other than core funder)

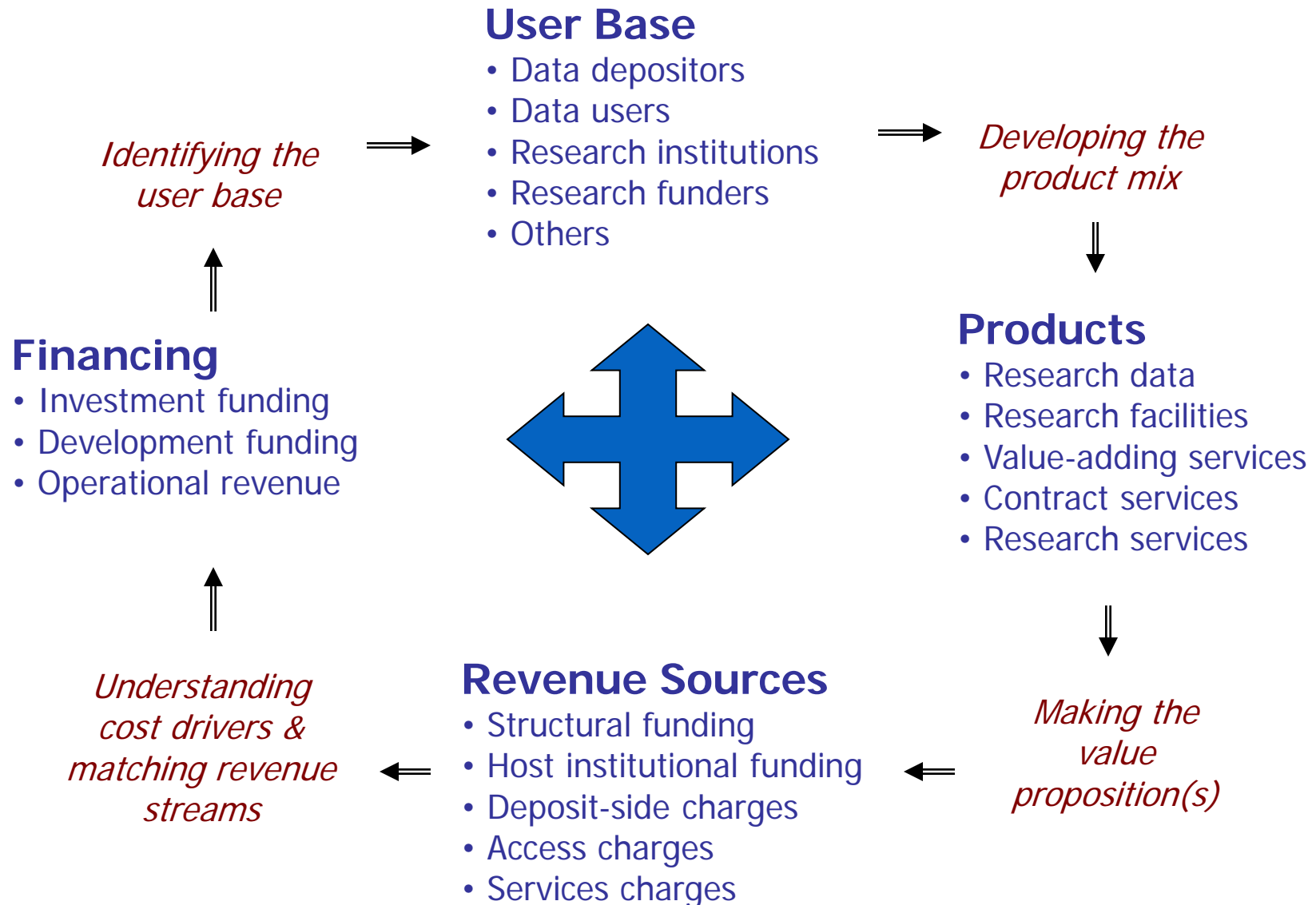


Sustainable business models for data repositories

- Continuation of the work under the umbrella of OECD/GSF
 - Around 50 interviews in total
 - Thorough economic analysis
 - Cost optimization
- 
- Stakeholder workshops
 - Presentation of report and stakeholder recommendations at RDA Plenary Montreal
 - Expected OECD publication end of 2017

<https://www.innovationpolicyplatform.org/open-data-science-oecd-project>

Elements of a Business Model for Data Repositories



Takeaways for the e-IRG LTP Guidelines

In order to realise the long-term preservation of data we need:

- FAIR data in TDRs
- Global network of TDRs
- Certification (at least CTS) to create trust
- Economic/organisational sustainability to enable long-term data accessibility



Thank you for listening

ingrid.dillo@dans.knaw.nl

www.dans.knaw.nl

