PARSE insight

Permanent Access to the Records of Science in Europe

What we learnt from PARSE.Insight

PARSE.Insight Symposium:

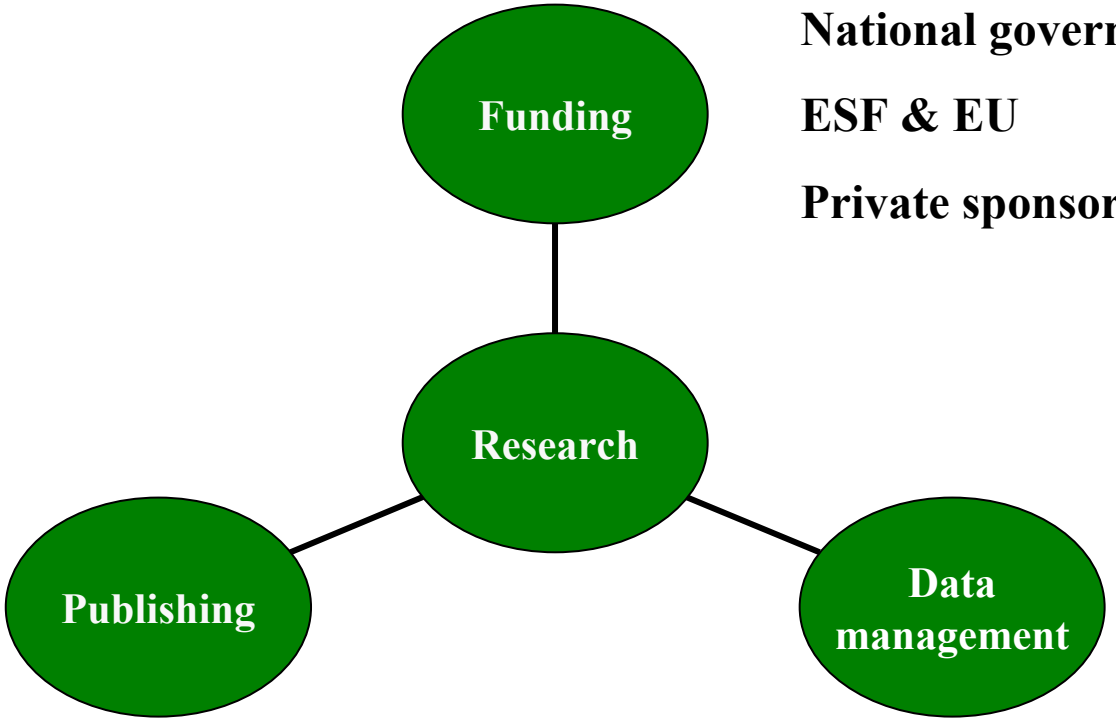
carrying the flame forward

Paris, 25th of June 2010

Jeffrey van der Hoeven, Salvatore Mele, Veronica Guidetti, Sabine Schrimpf

e-infrastructure
SEVENTH FRAMEWORK PROGRAMME

Our view on research



National governments

ESF & EU

Private sponsorships

Publishers

Aggregators

Online collaboratories

Nat. Libraries

Res. Libraries

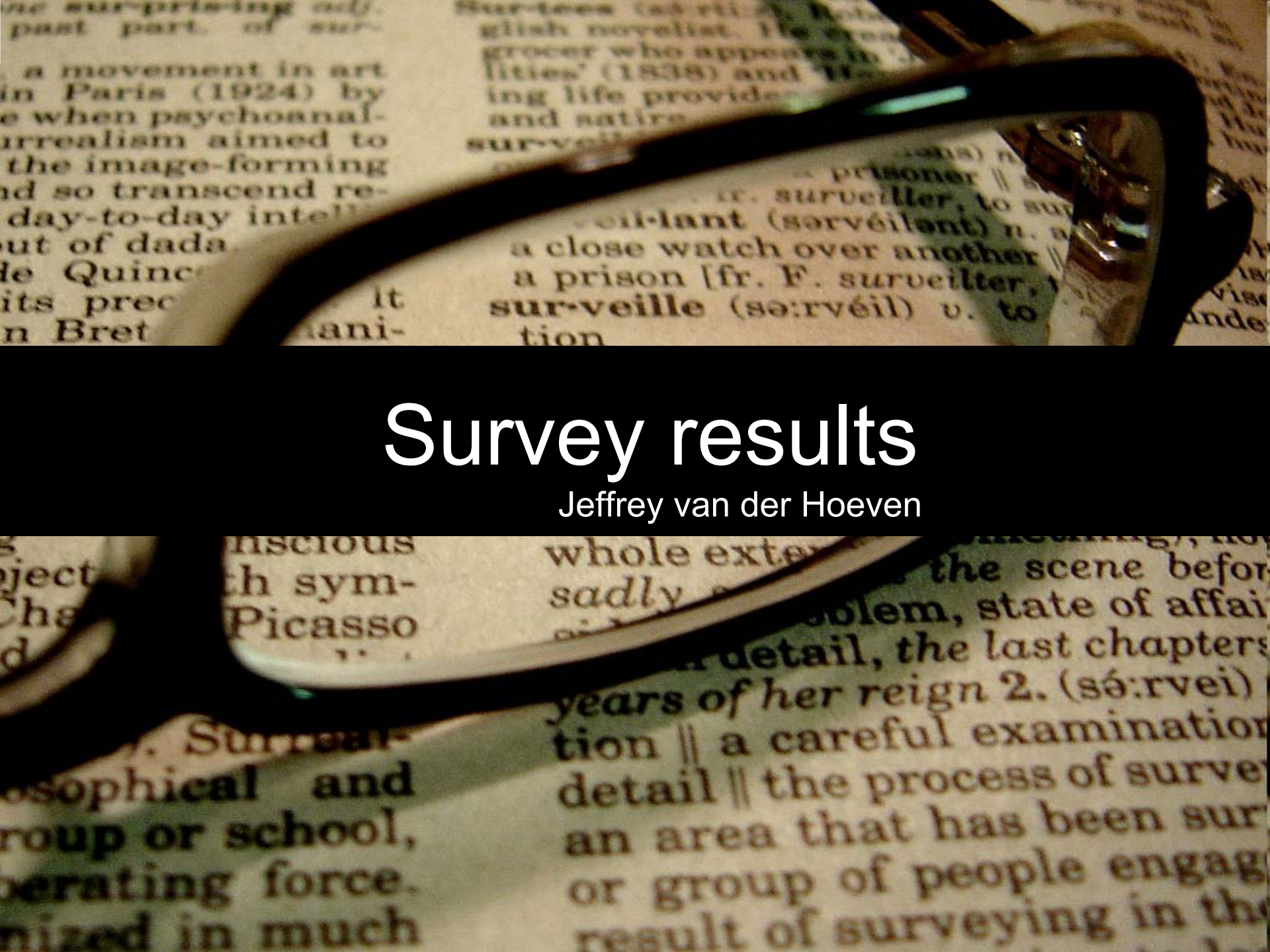
Archives

Data centers

Profit orgs

What do we have?

- Lots of survey results (over 2,000)
- Case studies:
  - High Energy Physics (investigated by CERN)
  - Earth Observations (investigated by ESA)
  - Social Sciences & Humanities, split into Book Studies (by DNB) and Psycholinguistics (by MPG & UGOE)
- In-depth interviews with funders
- Insights & recommendations for EC



Survey results

Jeffrey van der Hoeven

Main results from surveys

Overall:

- Agree on threats to and reasons for preservation of digital research data.
- Acknowledge the need for an science data infrastructure.
- Coherent approach for developing this infrastructure is missing. Unclearness about who should do what.
- National governments should pay the bill.
- Center of Expertise would be welcome.

Main results from surveys (2)

Researchers:

- Huge variety of data formats are in use. Lack of standardisation.
- Data is often open for their own research group, but less outside of that.
- They are reluctant to share data due to legal issues and misunderstandance of the data by others.
- Data often stored locally (own PC, server). They are not familiar with institutional repositories or archives.
- Guidance (training, tools) is needed.

Main results from surveys (3)

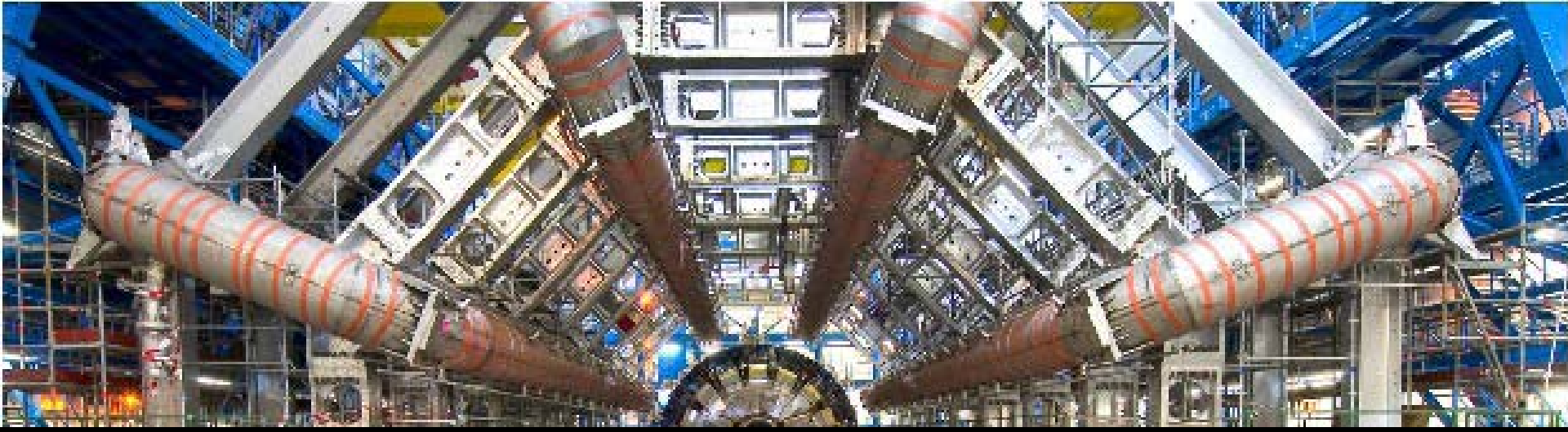
Data managers:

- Amount of data archives vary greatly per discipline and country.
- Majority of archives has a preservation policy.
- Mostly preserved: theses, journals and illustrative material in standard image and office formats.
- Other types of data (e.g. software, source code) is less covered.
- Data volumes grow explosively, shifting from TB to PB.
- Data linking is upcoming. About half of the archives say users can link to their data in journals.
- Current tools and infrastructure is not sufficient for upcoming challenges in preservation of data.

Main results from surveys (4)

Publishers:

- Large (71%) and small (57%) publishers acknowledge that authors can submit underlying research data with their publication. That covers 94% of journals.
- 84% of large and 55% of small publishers do have a policy for preservation of publications. 93% of journals.
- About 70% has no policy for data preservation.
- Most of small publishers have no preservation strategy in place. Large publishers often have arrangement outsourcing for this and/or apply normalisation/migration.
- Future scenarios: hybrid model of subscription-based and open-access journals, while subscription model remains dominant is opted most.
- Publication itself will become more diverse (multimedia)



High Energy Physics

Salvatore Mele

Ref: CERN

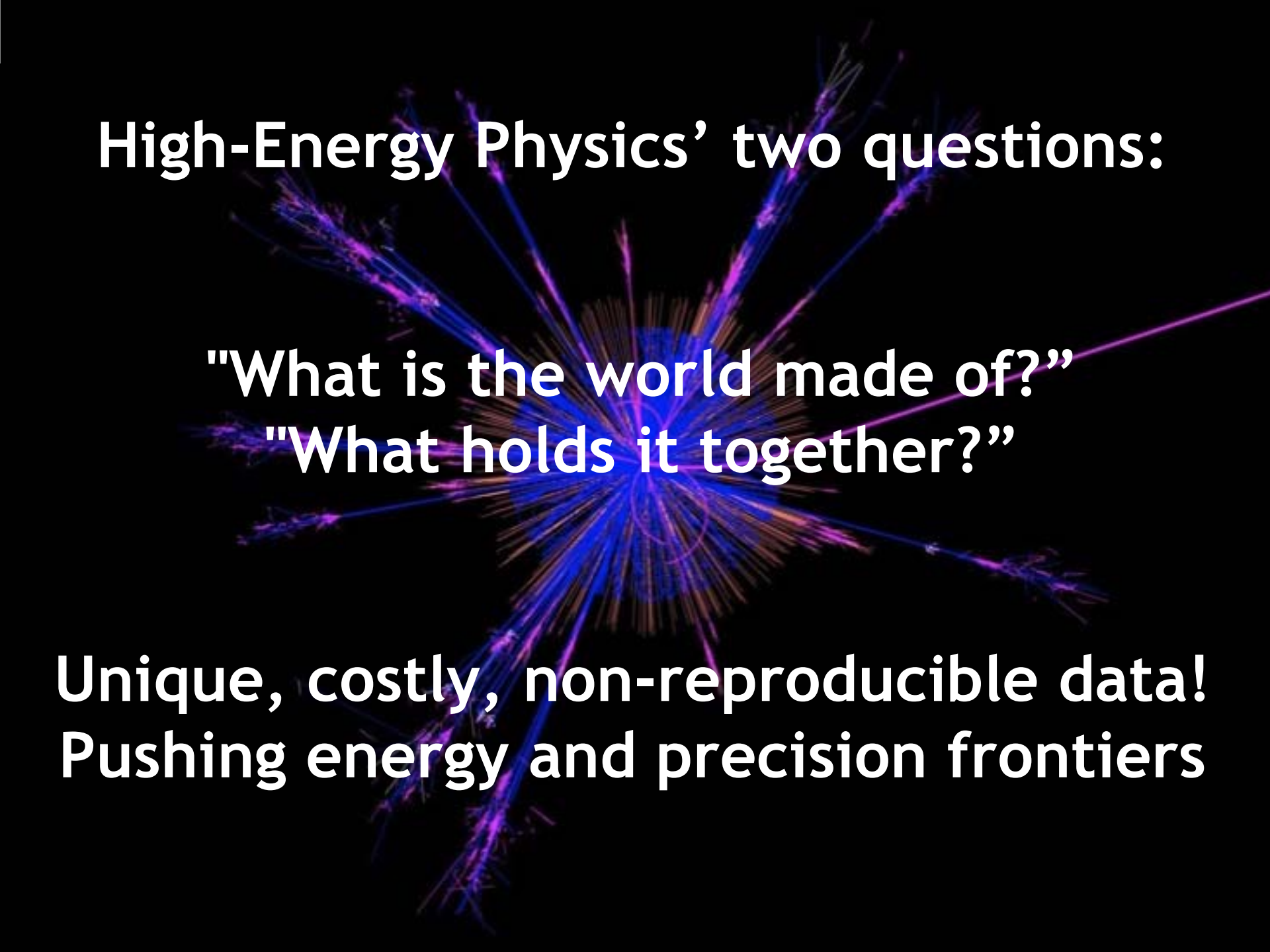
e-infrastructure
SEVENTH FRAMEWORK PROGRAMME

Data preservation, re-use and (open) access. A case study in High-Energy Physics

Highlights from two years of surveys, interviews, workshops

Andre Holzner (CERN/UCSB), Peter Igo-Kemenes (Gjøvik/CERN), Salvatore Mele (CERN)

PARSE.Insight Final Event
Paris, June 25th, 2010

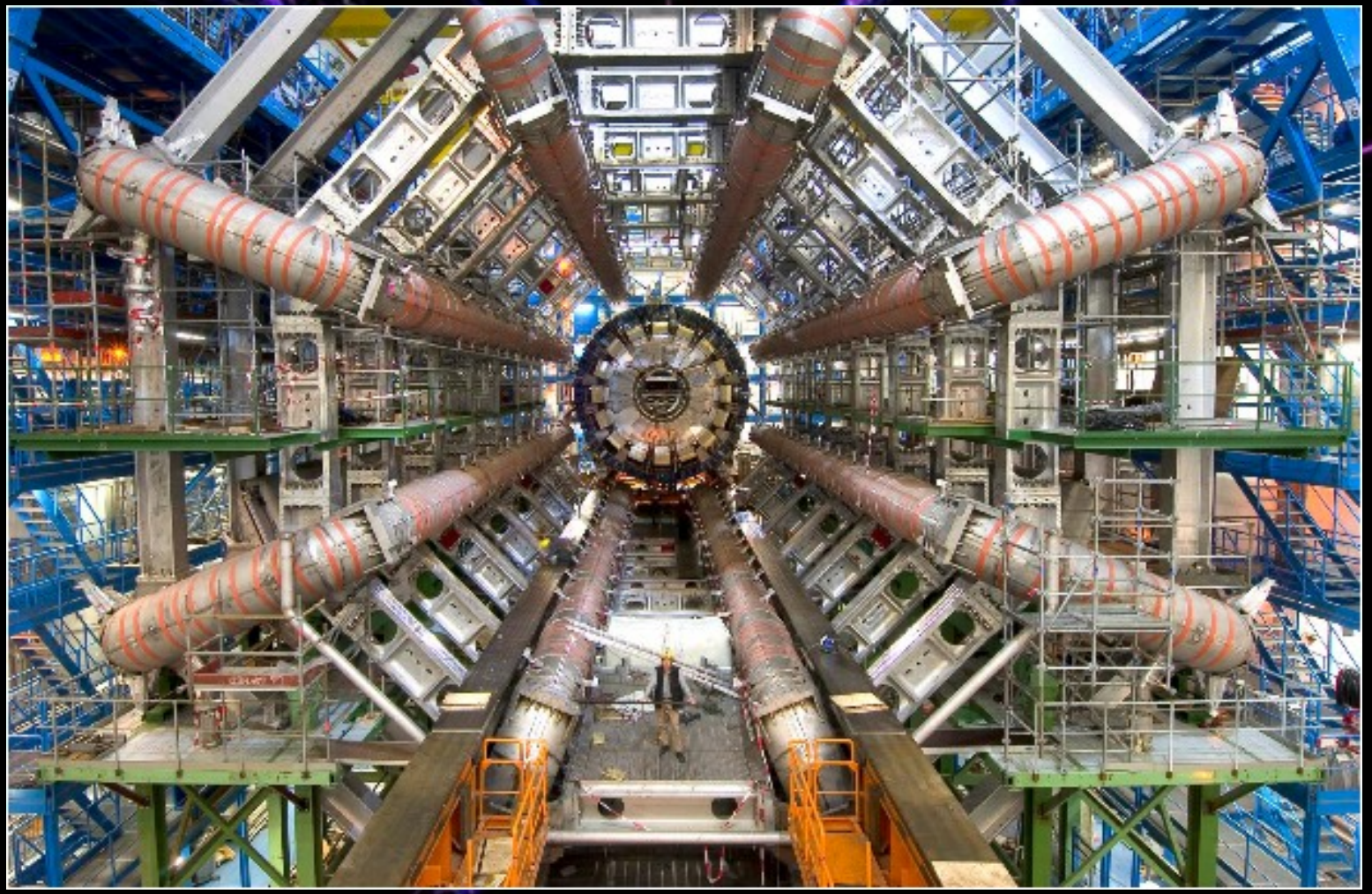


High-Energy Physics' two questions:

"What is the world made of?"
"What holds it together?"

Unique, costly, non-reproducible data!
Pushing energy and precision frontiers

~15'000 High Energy Physics (HEP) scientists smash stuff at the speed of light to produce new stuff (and 15PB of data/year!)



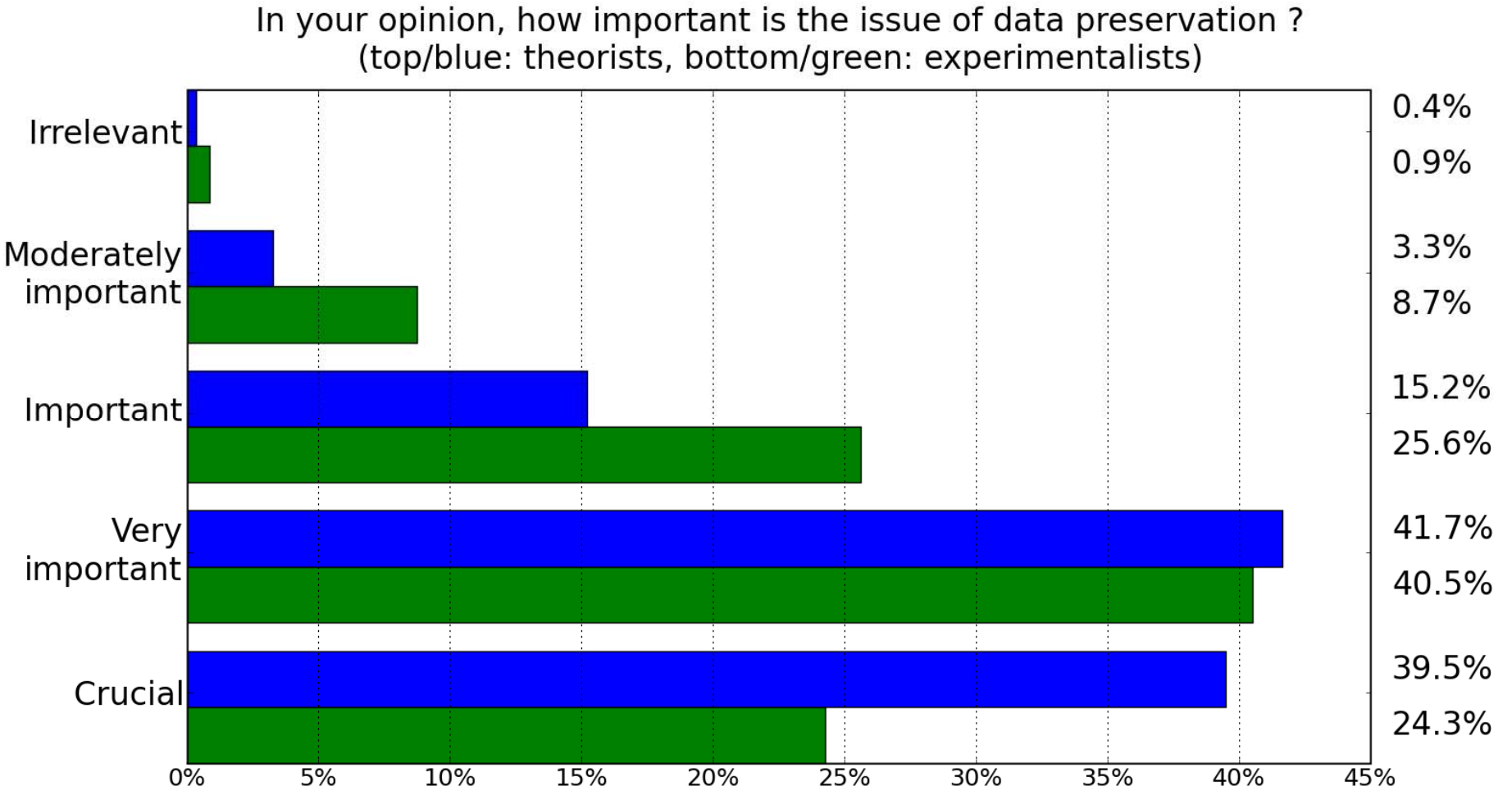
~15'000 HEP theorists scratch their heads to make sense of all that stuff and then some more (and might later want to look into the data)

Who cares ?

PARSE insight

- Same researchers who took data, after the closure of the facility (~1 year, ~10 years)
- Researchers at similar facilities at same time (~1 day, ~1 week, ~1 month, ~1 year)
- Researchers of the future (~20 years)
- Re-interpretation by theoretical physicists (~1 month, ~1 year, ~10 years)
- Theoretical physicists testing future ideas (~1 year, ~10 years, ~20 years)

They care !



“Very important” + “Crucial” = 69%

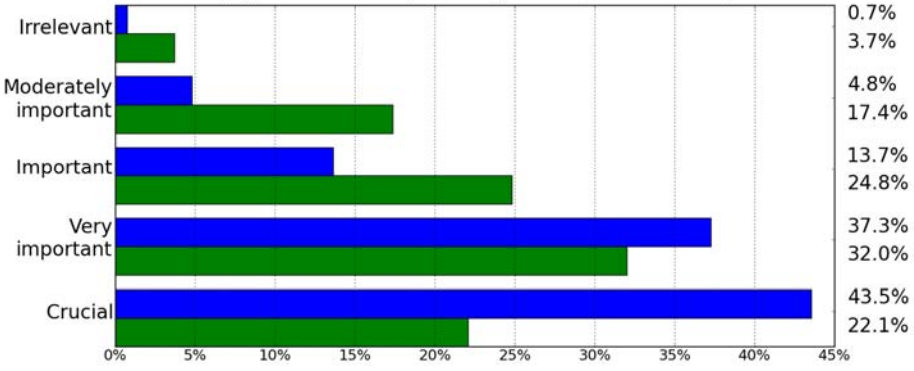
[Preservation] is a step toward truly open access. The fact, that [HEP] data is closed is a historic accident. Science does not benefit from it being closed.

Respondent # 254

Drivers: accelerate science!

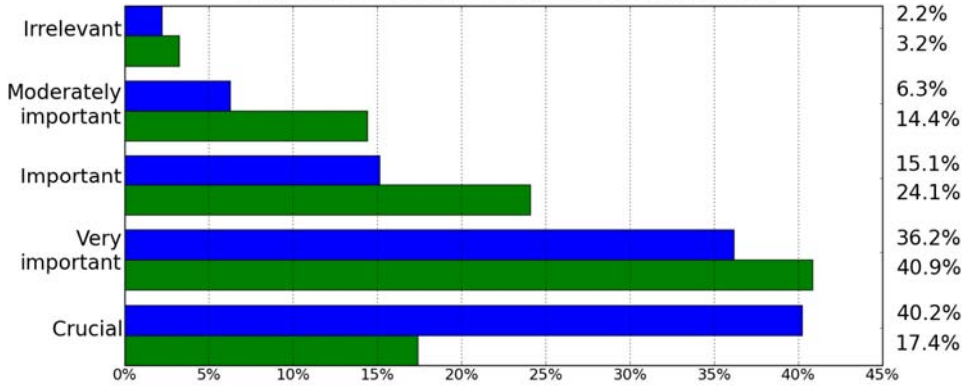
Future independent checks

b) Preserved data would allow future independent verification of results
(top/blue: theorists, bottom/green: experimentalists)



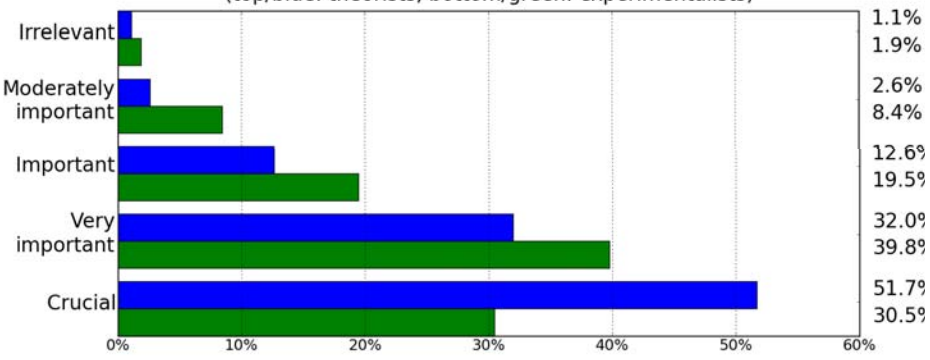
Re-use with future data

c) Preserved data could be reused in combination with future data
(top/blue: theorists, bottom/green: experimentalists)



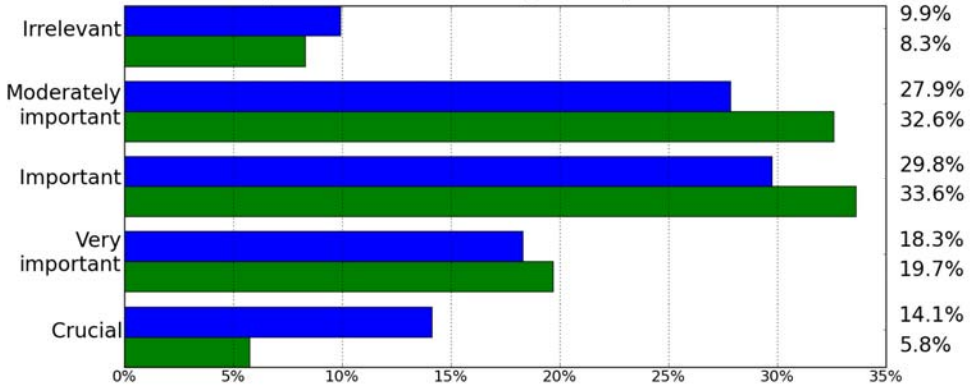
Re-analyse (future theories)

d) Preserved data could be re-analysed in the light of new theories
or experimental results
(top/blue: theorists, bottom/green: experimentalists)



Teaching and outreach

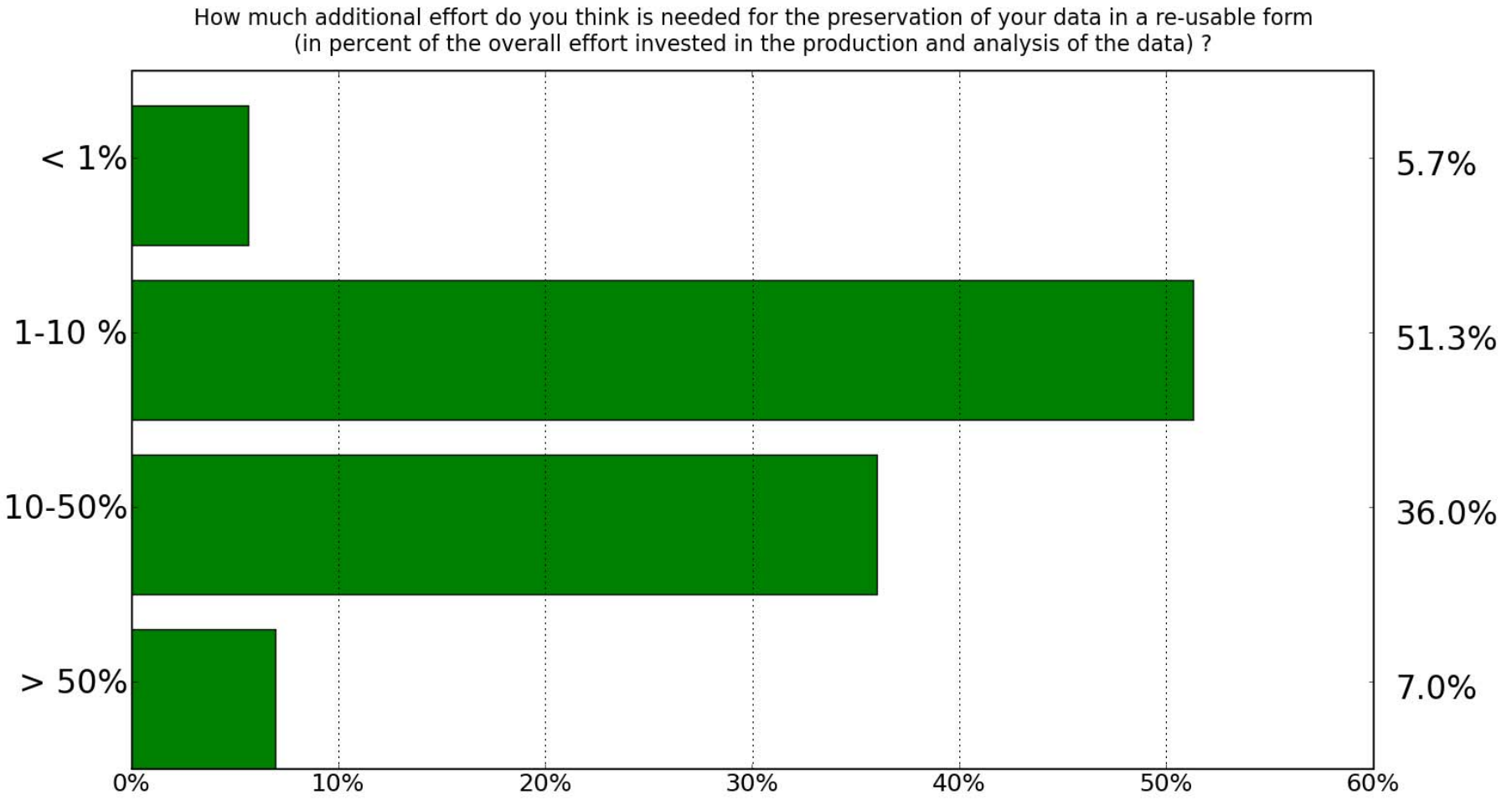
e) Preserved data could be used for teaching and outreach
(top/blue: theorists, bottom/green: experimentalists)



PARSE insight
Permanent Access to the Records of Science in Europe

Respondent # 987

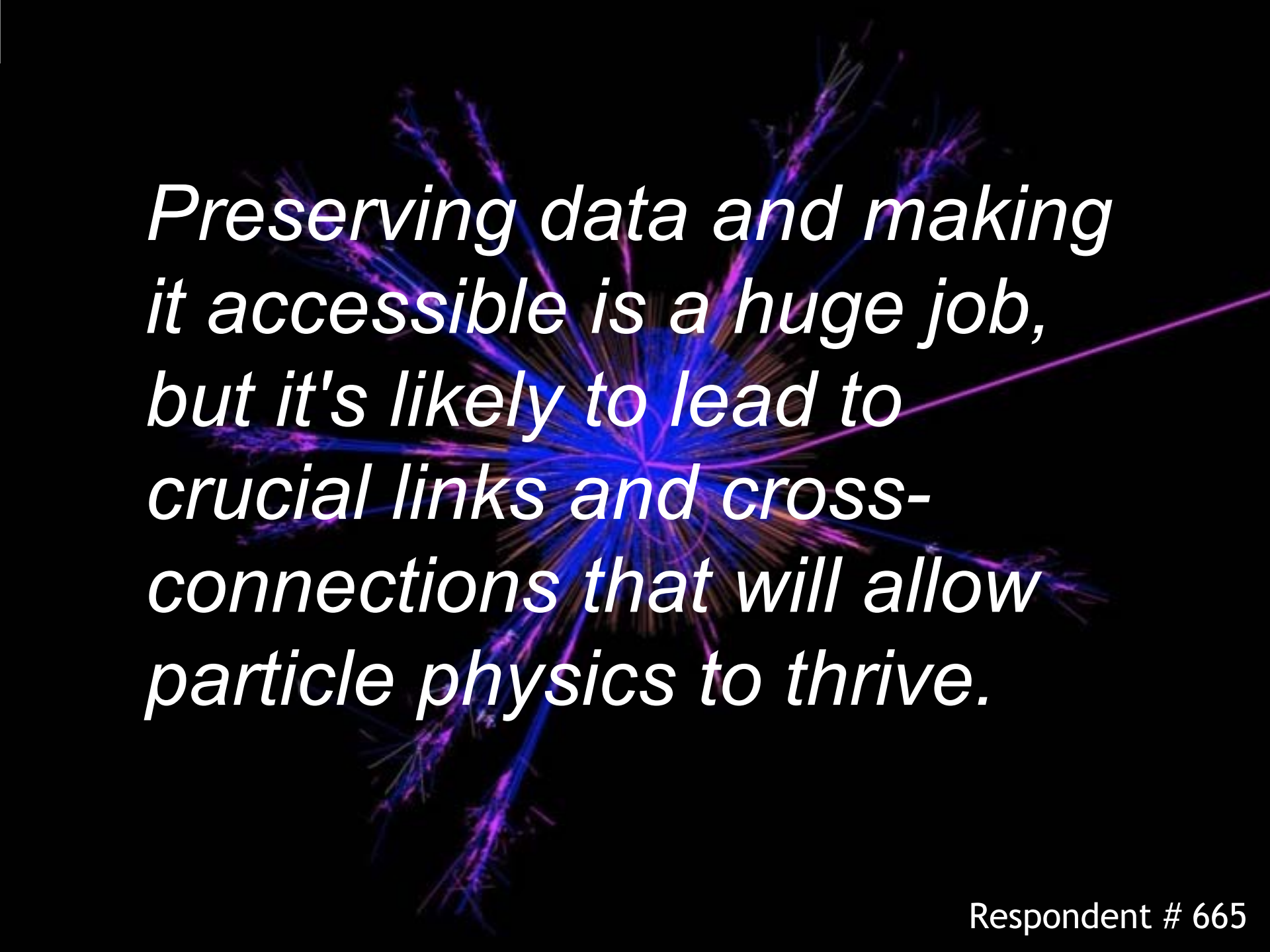
Barriers: high human and financial cost



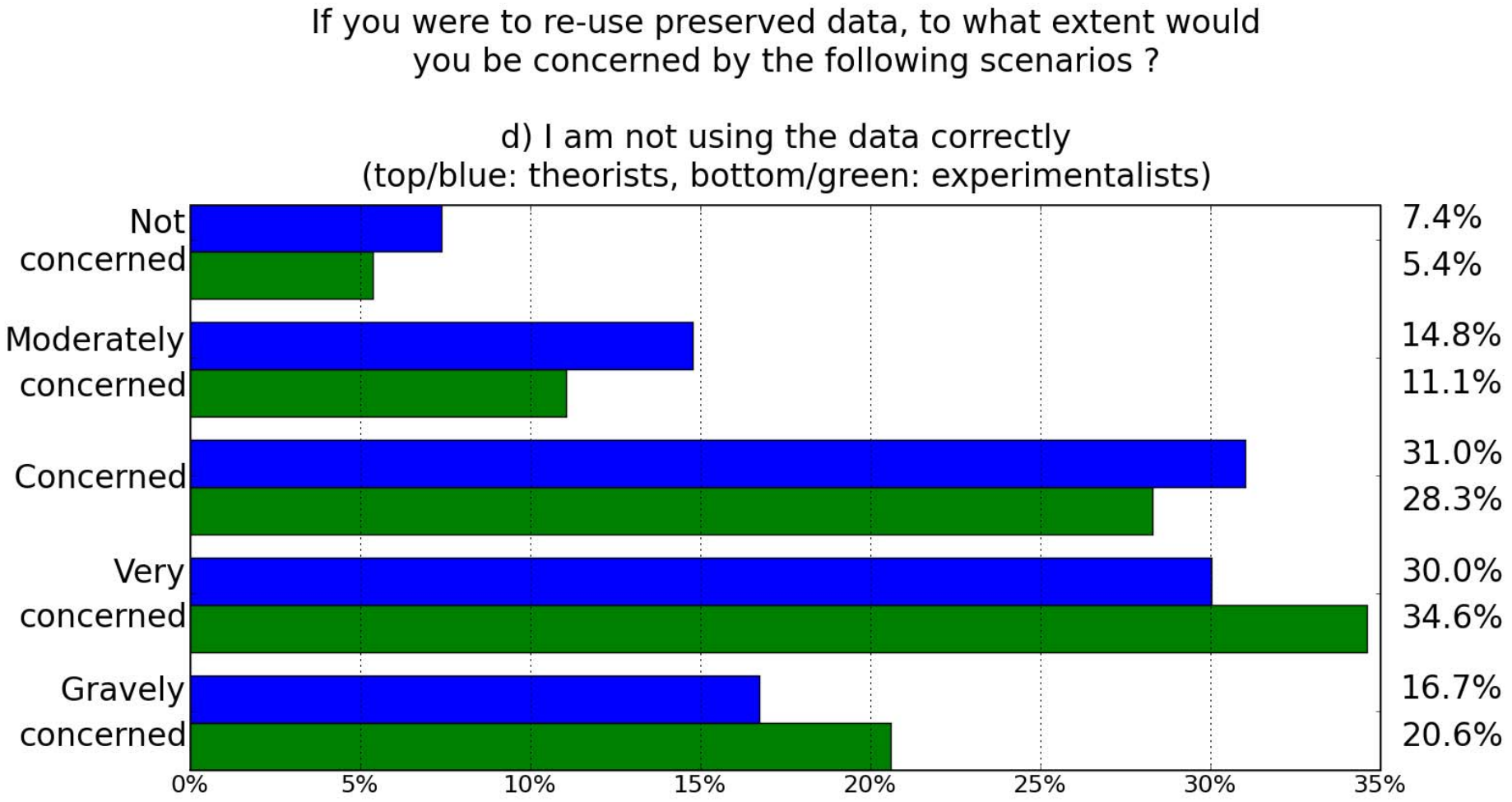
Additional effort on top of taking data…

…what sums up to thousands of person-years!

PARSE insight
Permanent Access to the Records of Science in Europe
CERN
e-infrastructure
SEVENTH FRAMEWORK PROGRAMME


Preserving data and making it accessible is a huge job, but it's likely to lead to crucial links and cross-connections that will allow particle physics to thrive.
Respondent # 665

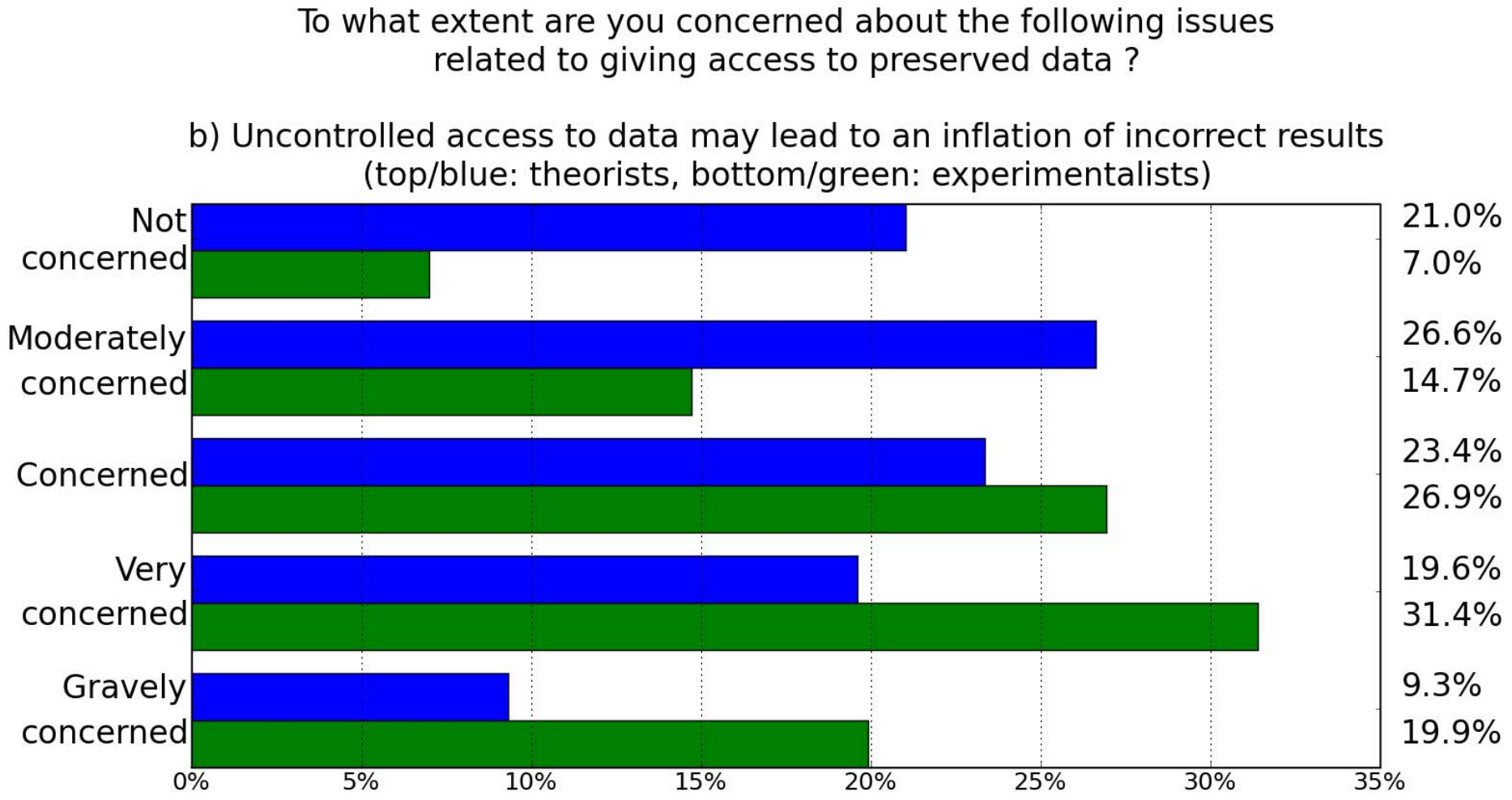
Barriers: complexity and documentation

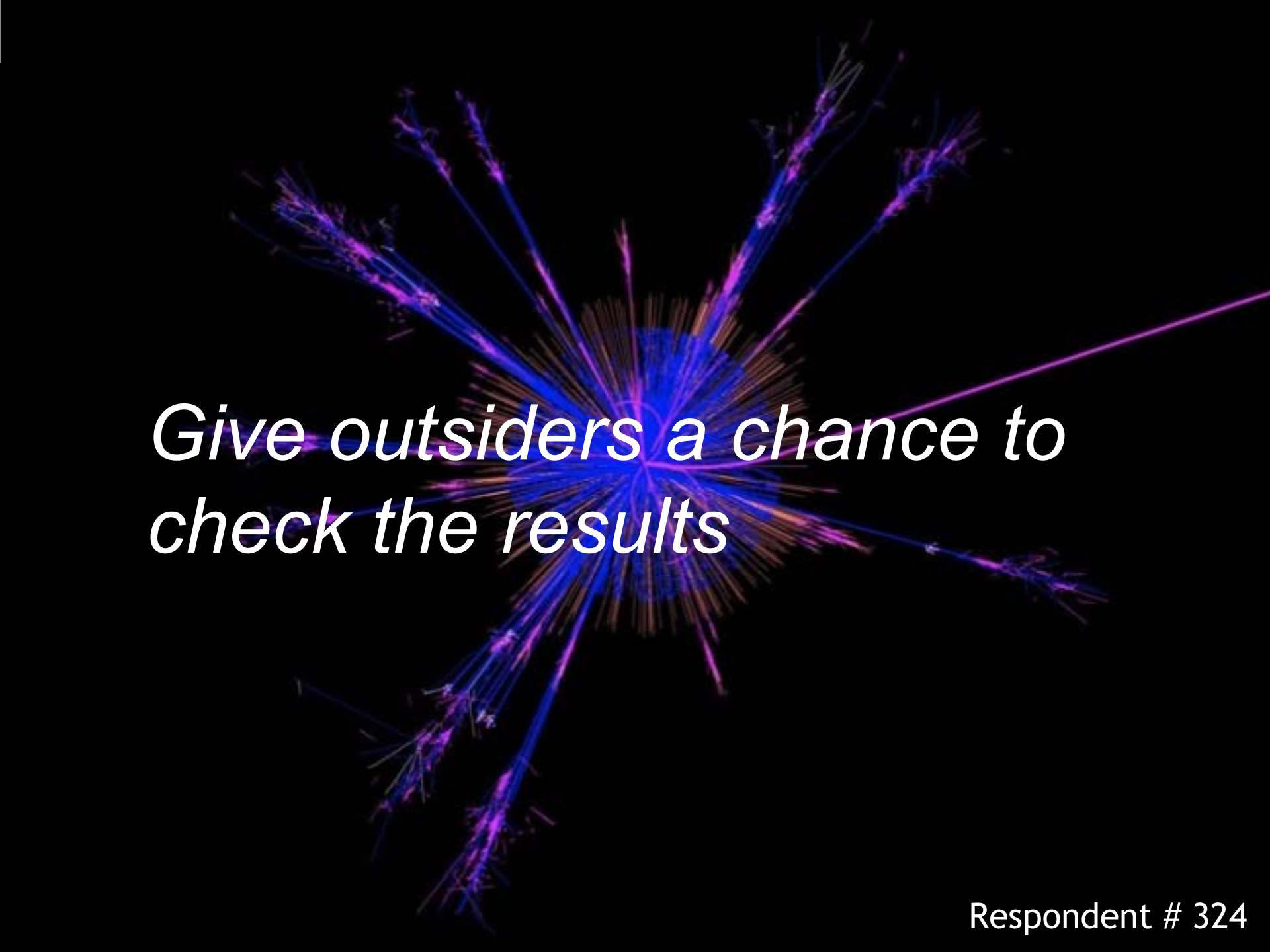


Documentation, documentation, documentation; that's the real bottle-neck, you'd need collaborators willing to explain the real limits of how the data can be used to reach scientifically valid conclusions

Respondent # 248

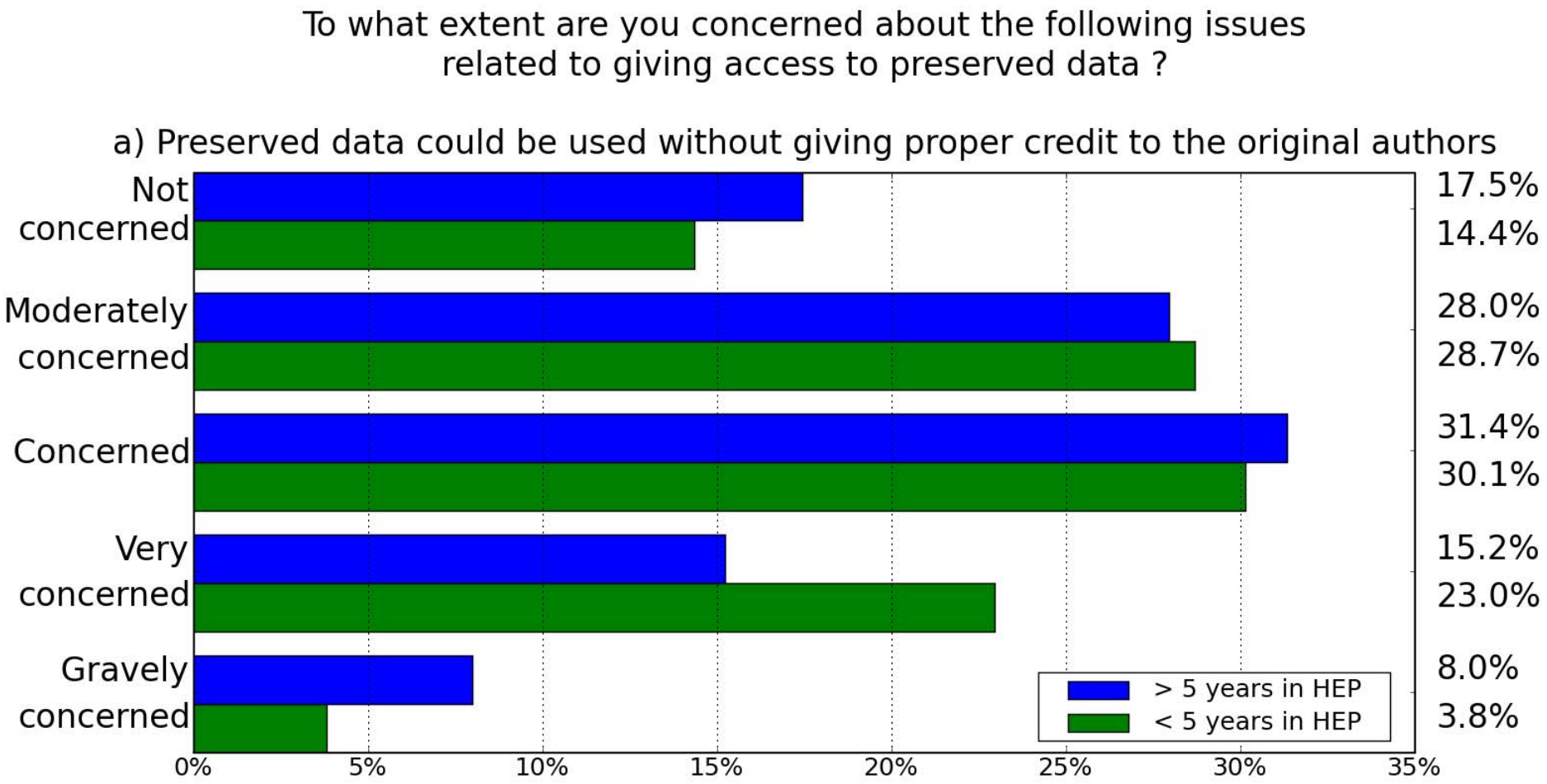
Barriers: is it just producing inflation/noise?





Respondent # 324

Barriers: lack of recognition

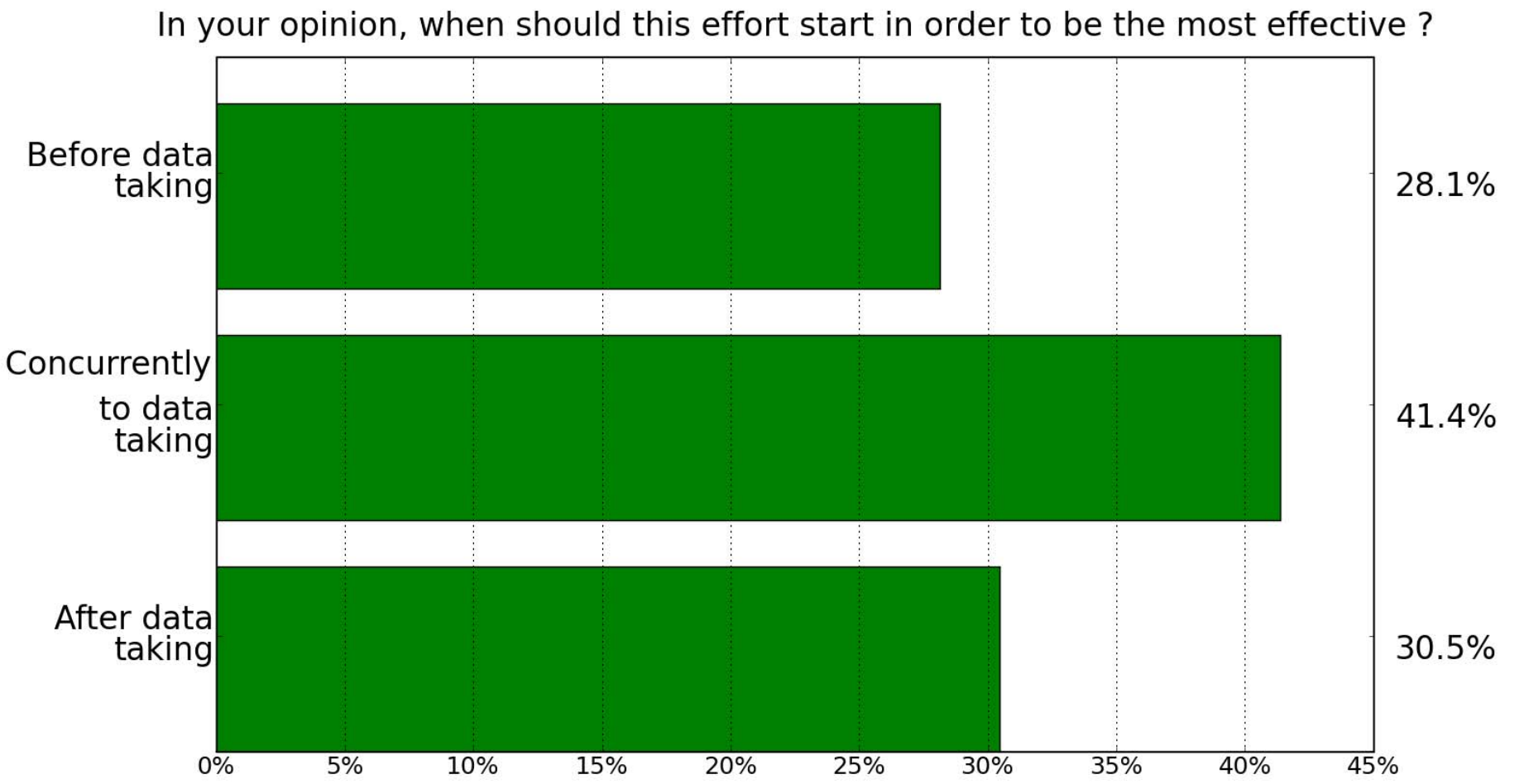


Q: When would you invest effort into preservation ?

A: *Whenever I have time to do so. This would not be a priority because there would be no recognition for the large effort involved*

Respondent # 998

The gap: conflict in effort timing!



Note: ALL non-reproducible data already taken!

Respondent # 137

Why should we preserve data?

- Data are unique, costly, irreproducible
- High human and financial investment
- There is more in data than it appears
- Important data is practically already lost

What is in place?

- Little beyond strings of 0's and 1's
- Disperse, proprietary, repositories
- No large-scale infrastructure
- Few links between data and publications

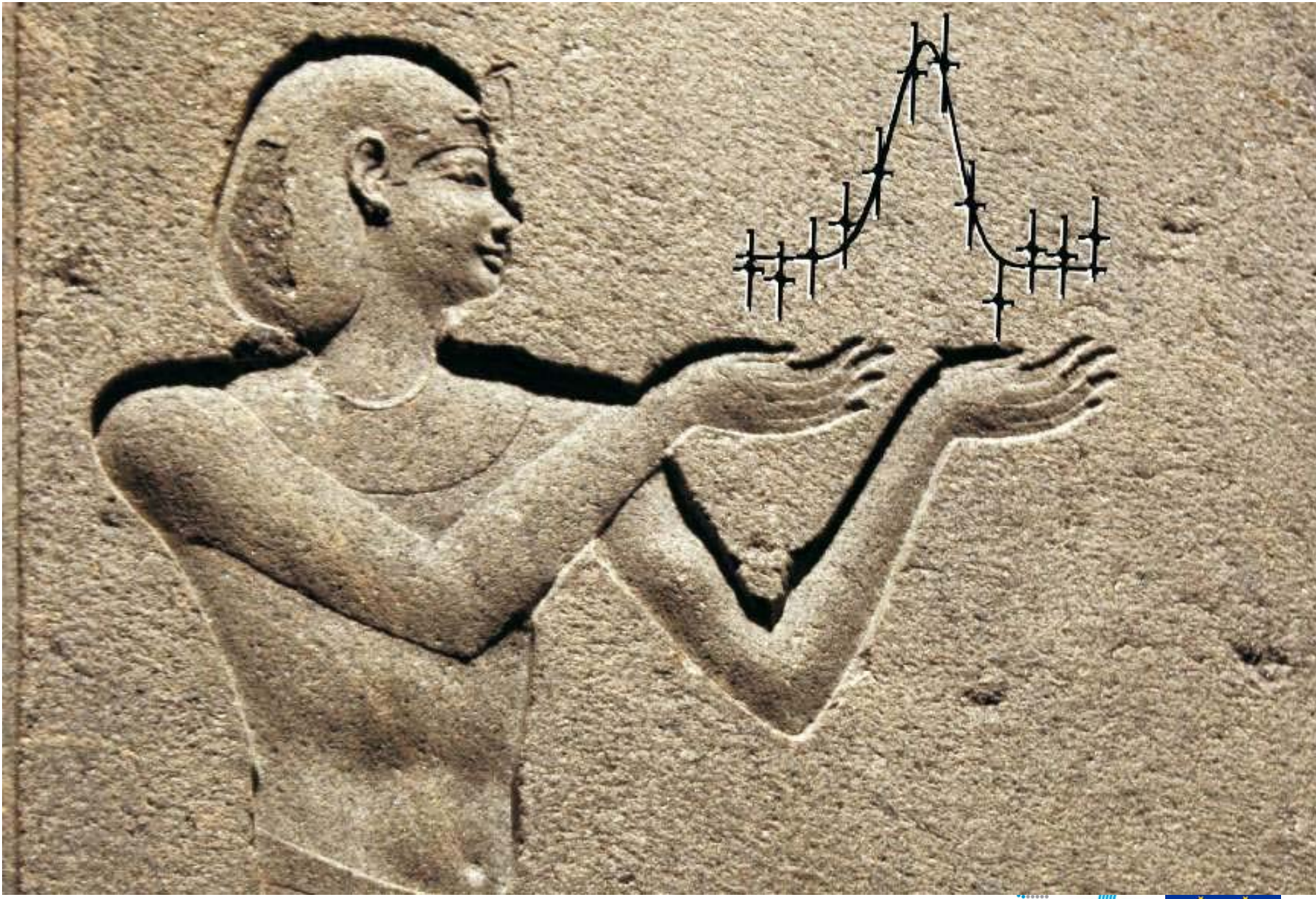
Which are the needs?

What are the risks if nothing is done?

- 0's and 1's become scientifically useless
- Loss of data of last-of-its-kind facilities
- Missed unique scientific opportunities

What is being done?

- PARSE.Insight rose awareness
- DPHEP.org group is drawing a blueprint
- Design next-generation e-Infrastructure
- Aim to pragmatic bridges on the gaps
- Delicate exercise for governance/access
- Seeking funding agencies endorsement



Earth Observation

Veronica Guidetti

Ref: Buenos Aires (Argentina). Picture taken by ESA on 13-06-2008
http://earth.esa.int/satelliteimages/1922/Buenos_Aires_MER_FRS_20080613_32872_cut_or.jpg

e-infrastructure
SEVENTH FRAMEWORK PROGRAMME

Short overview of discipline

Used to monitor the status of natural and built environments

Data acquired from space are a world-wide asset and their preservation has to be considered as a long term, challenging, responsibility for the Space Agencies and the data owners

Why should we preserve our data?

EO archives and datasets are invaluable:

- Analysing the state of the Earth, its environment and its variability over time requires a very large number of observations;
- It is impossible to go back in time and resample environmental data, therefore global and complete measurements need to be performed;
- The value of an environmental data set is impossible to estimate and it is impossible to foresee its potential future uses

Why should we preserve our data?

e.g. the international activity in support of climate change

Earth observations are of unique value to understanding climate change and to develop an appropriate response to global and regional climate change

Critical parameters are derived by processing global historical time-series, re-analysing archived data using models and data simulation systems

This requires periodical re-processing mission datasets

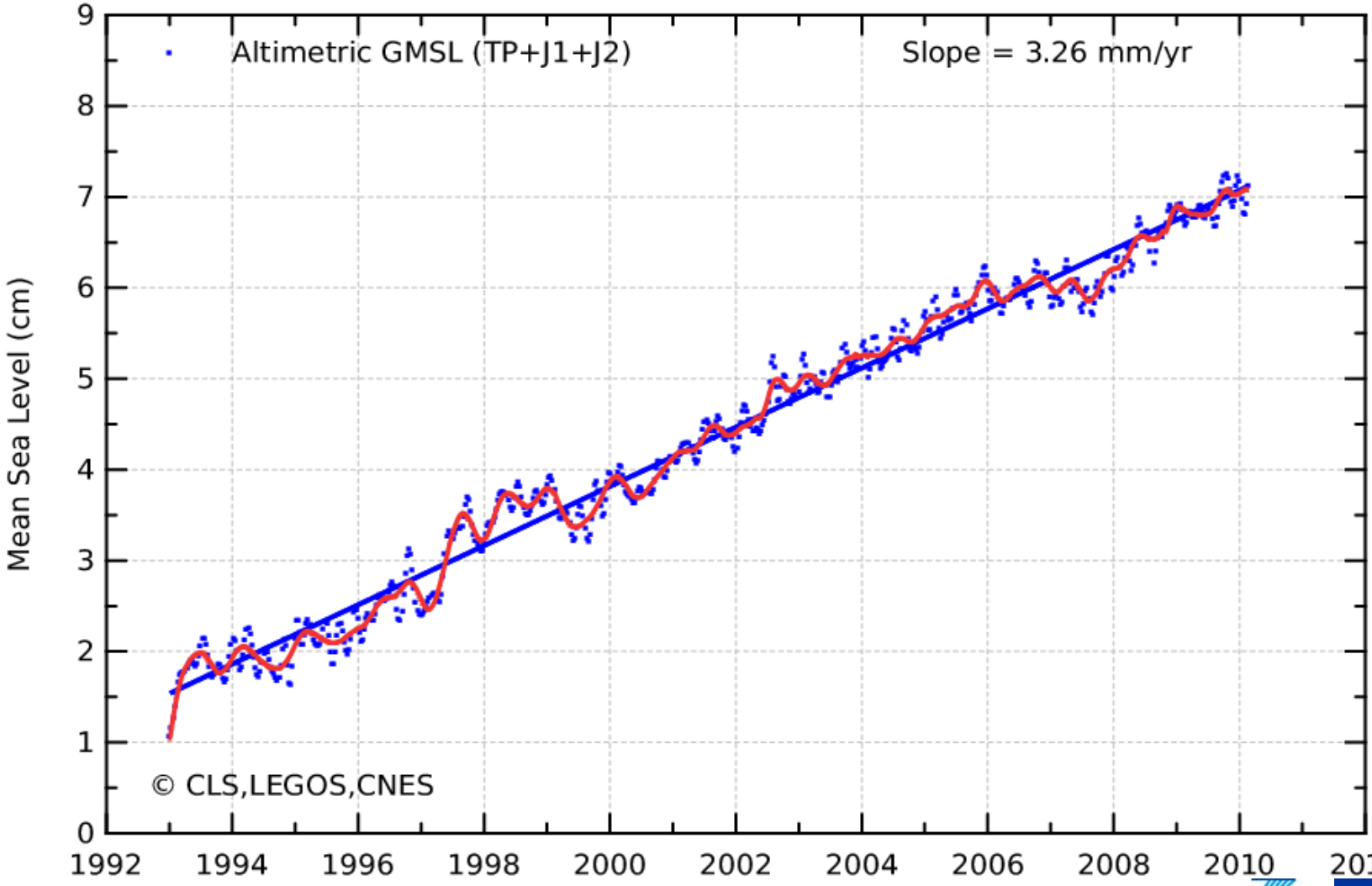
Mission dataset re-processing

- Issue: mission data set not fully homogenous over the entire mission lifetime (different algorithmic evolutions of ground processing chains…….)

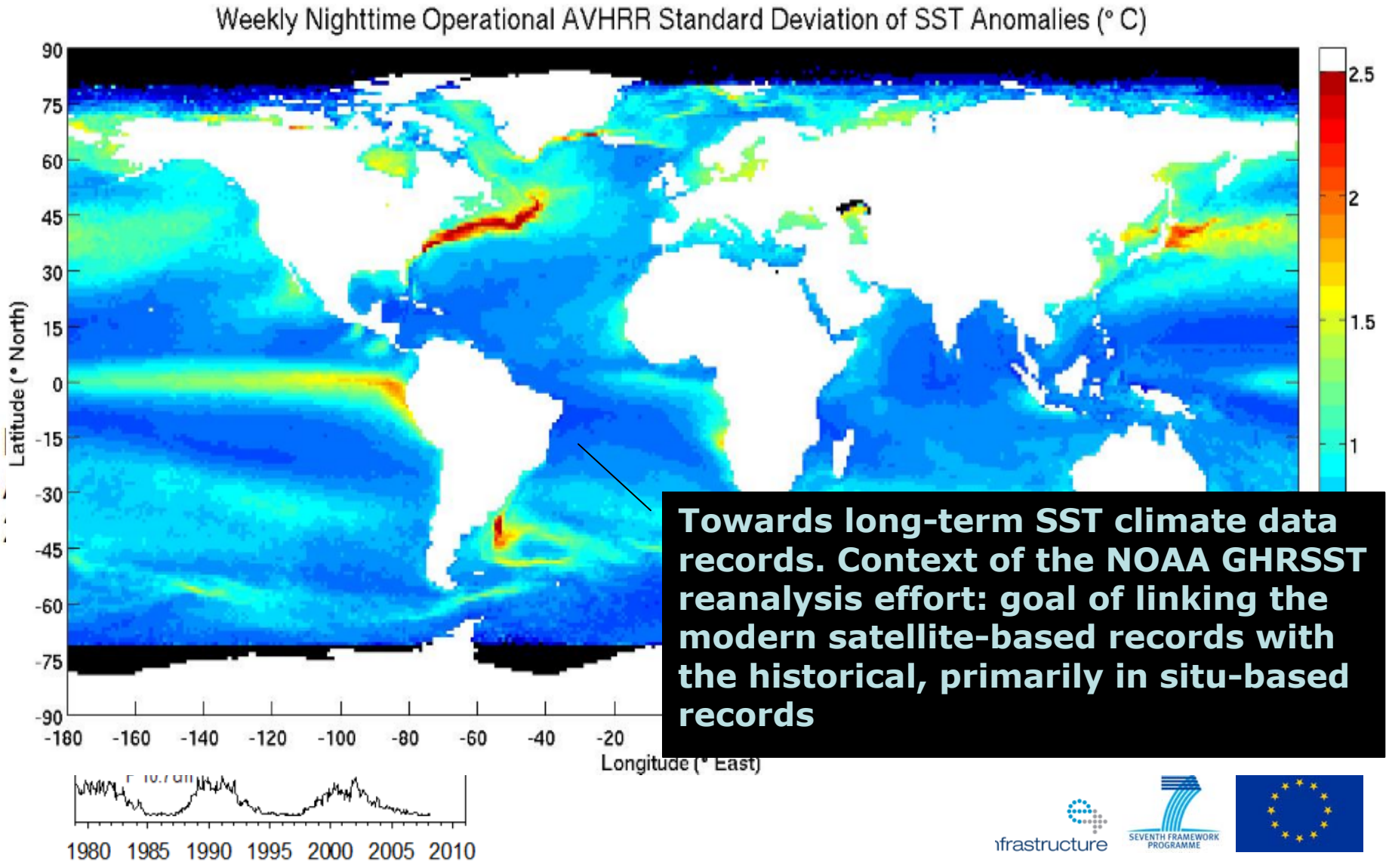
- Target: to reach an homogeneous long-term series with higher accuracy than presently available. This would allow for the exploitation of the whole potential of the archived mission data set. Application example: calculation and study of the Global Mean Sea Level…

- Solution: to issue the re-processing of the mission
- Pre-requisite: the dataset must be traceable, accessible and well known!

Mission dataset re-processing

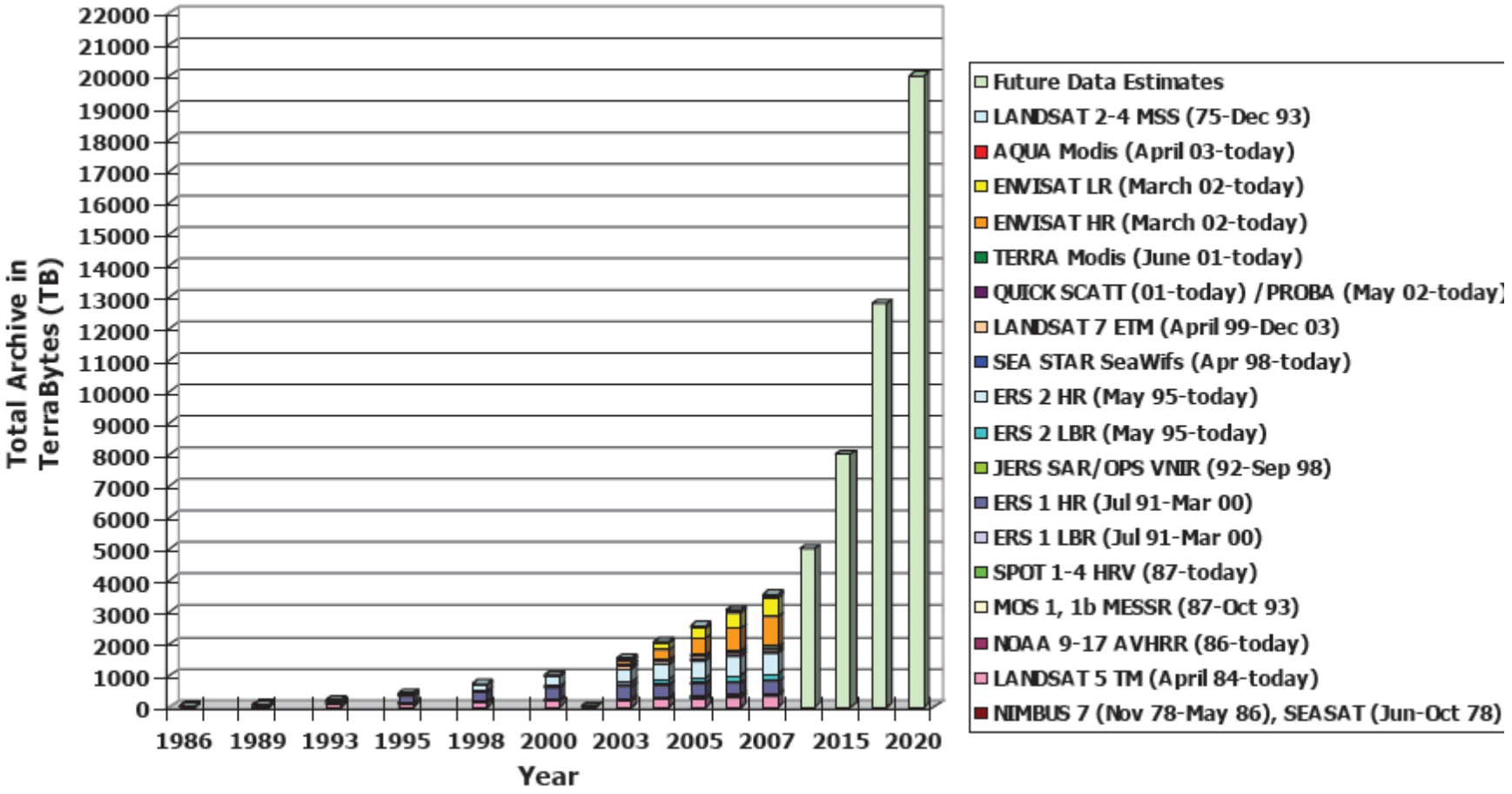


Further applications



Towards long-term SST climate data records. Context of the NOAA GHRSST reanalysis effort: goal of linking the modern satellite-based records with the historical, primarily in situ-based records

What are the needs?



Current infrastructure

The relentless growth of EO archives' volume requires upgrades to the capacity of storage media, data management and cataloguing, robotics, data security and computing systems, including adaptation to new standards…

Despite the increasing capacity, resources are limited, and so there are restrictions on the volume and quality of data that can be archived…

Envisaged infrastructure

- Based on very high performance computing and distributed architecture (scalability)
- It should enable full re-processing of global data on timescales of weeks instead of years, dramatically reducing the time for production and delivery of EO data products
- It should allow easy plug-in of newly developed algorithms and easy access to auxiliary data and information (discovery)

Actions taken

Related ESA programmes and initiatives include CCI, LTDP, APA and involvement in EC FP7 actions…

Criteria and conclusions

Benefit

- Following the operation of upcoming missions, data transfer over high capacity networks will be necessary between facilities' archives to enable re-processing of long time-series data, as well as access to TBs of on-line products and in-situ data

- Beyond enhanced means for re-processing capacity and data distribution, the infrastructure must ensure easy access to information, possibly based on the Web2.0 paradigm

Worst case scenario

- The lack of availability or inaccessibility of/to spatial data and information would prevent e.g. studies of changes in climate, pollution, ice cap coverage, area of glaciers, surface temperature, the effect of which would be dramatic for the whole human race

Sustainability

- A sustainable infrastructure can be reached only by passing from a demonstration and prototyping phase, including a number of projects and dedicated initiatives, to an operational one

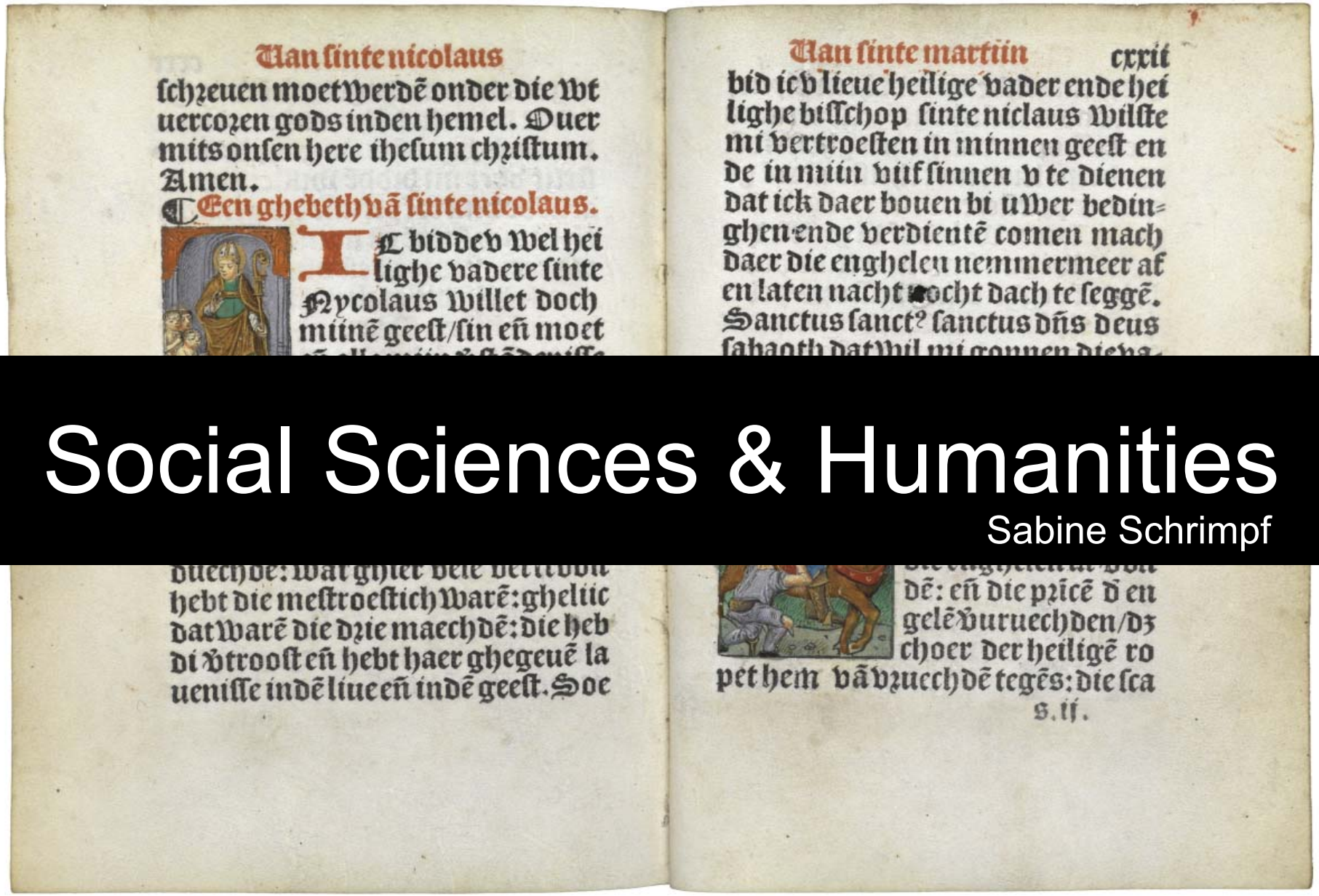
- It is important that other disciplines be enabled to benefit from such an infrastructure as well

Resources

- A trustworthy sustainable infrastructure needs funding
- Collaboration, technology sharing, common implementations for various disciplines, results exchange in the field of digital data preservation should represent continuous resources, to be provided along all the infrastructure building life-cycle
- Linking to significant data user communities, PARSE.Insight and its roadmap constitute a first step forward in the direction of building the infrastructure required

Community perspective

- The scientific community need to access historical environmental data
- The community aim at a more active involvement in the process not only via reporting experiences and suggestions, but as customers of a global information system
- Users are aware and informed about current infrastructures' limitations to enable data availability and accessibility, and ask for timely and effective solutions
- The global community requires similar services: new programs together with evolving user needs impose the implementation of similar services both for historical and current missions
- The whole focus for an enhanced infrastructure is on end-to-end (operational) systems, where all aspects relating to digital preservation are properly considered



Social Sciences & Humanities

Sabine Schrimpf

Ref: Medieval book pressed by Wolfgang Hopyl in 10 Sept 1500 (part of KB collection)

http://www.kb.nl/pr/pers/fotos/getijdenboek/index.html

e-infrastructure

SEVENTH FRAMEWORK PROGRAMME

Short overview of discipline I

Why Social Sciences and Humanities?

- We assumed that their data usage and e-infrastructure needs differ from those of the hard science communities

Why Psycholinguistics and Book Studies?

- Psycholinguistics: data is created during experiments and/or observations
- Book studies: work with digitized objects provided by a third party
- Together, they represent the spectrum of Humanities and Social Sciences

Short overview of discipline II

Psycholinguistics

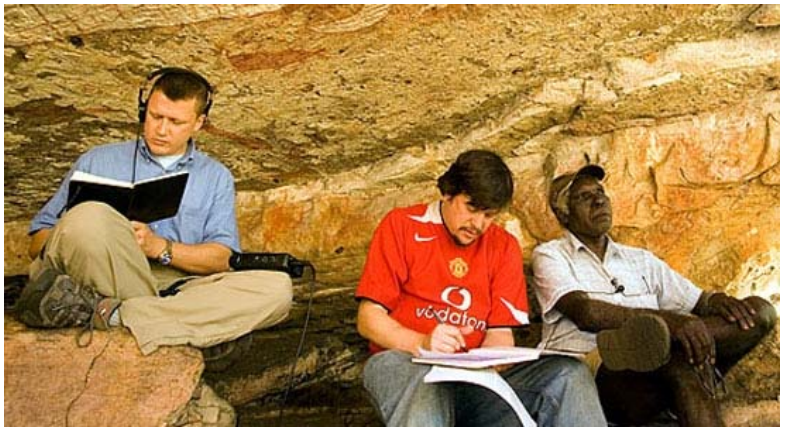
- Concerned with the relationship between the human mind and language as it examines the processes that occur in the brain while producing and perceiving both written and spoken discourse

“Book Studies”

- Developed in the 19th century as an ancillary science from the study of literature
- Media-Science, which takes part on the exploration of a cultural history of knowledge
- Traditional objects of research are ancient books, the invention of printing by Johannes Gutenberg as well as actual tendencies on the book market

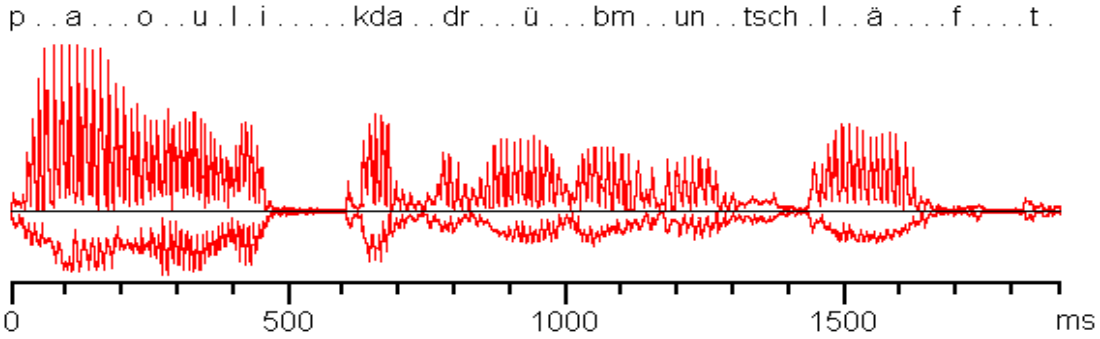
Why should we preserve the data?

Psycholinguistics: Unique primary data created during research workflow



Recordings of spoken language

Language production research



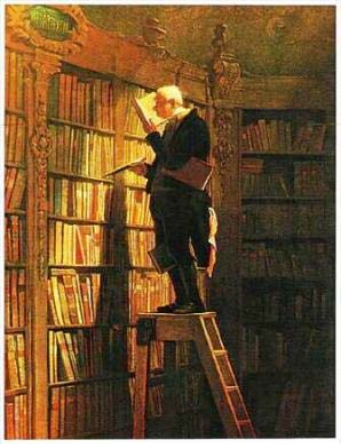
Analysis of speech patterns

Why should we preserve the data?

Book Studies: No clear, common understanding of "research data": Sources? Researchers' notes? Unpublished manuscripts?

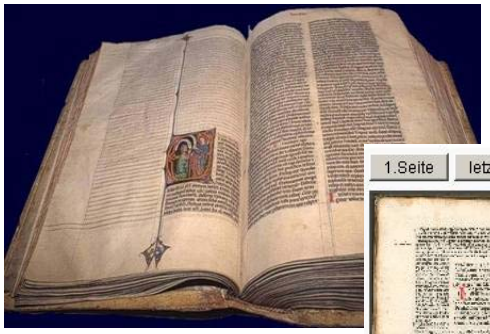
New field of action: Preservation of digitizations

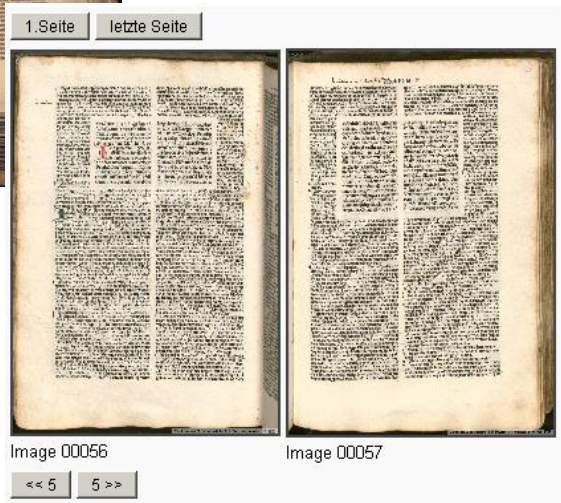
Research in an archive/library



Hermeneutical study of sources

- physically or online





Research in databases

What is in place?

What is already there in terms of general e-Infrastructure?

- PL: language archives, specialised databases (e.g. audio files, lexical corpora)
- BS: library catalogues, specialised databases (e.g. of manuscripts, early printed books), many digitization initiatives

What are the needs?

- Most respondents think that an international e-infrastructure would help guard against the threats to digital preservation
- Ideas range from centralized EU agency to completely decentralized e-infrastructure

Psycolinguistics

Book Studies

What are the needs?

Book Studies

- In the Book Studies, importance of cross-discipline infrastructures was emphazised (need to make use of materials from other disciplines)

Psycholinguistics

- Only 2% are satisfied by the status quo, while 59 % claim that the e-Infrastructure should be improved.
- Especially, training in data curation is missed

What are the risks if nothing is done?

Book Studies:

- Loss of unpulished research notes
  - Only 10% submit them to a digital archive, and 11% to an external web service
  - 74% were willing to use a digital archive (of their organisation), if there was one and 69% would make use of a discipline specific repository
- Obsolescence of databases that hold digitizations

Psycholinguistics:

- Existing infrastructure does not promote data exchange and cooperation between research groups
- Unique and irreproducible Data become unusable due to the rapid technological changes
- The “digital divide” between SSH and natural sciences may increase rather than be bridged

What measures of improvement are being taken?

- National Data Service Centres as focal points for improvements in the Social Sciences
- Federation of European Data Centres:
  - Council of European Social Science Data Archives (CESSDA)
- First large scale projects in the Humanities are under way: CLARIN, DARIAH…

What do funders think…

Jeffrey van der Hoeven

Quotes from funders

"The problem with floppy disks and the loss of NASA's records of the first moon landing are two of the most striking examples of what can happen when digital preservation is not taken care of properly."

Interview Dutch Ministery of Education, Culture & Science (OCW), 8 Feb 2010

Quotes from funders

"When no repositories exist, how would it be possible to enforce researchers to archive their research data?"

Interview Deutsche Forschungsgemeinschaft (DFG), 31 March 2010

Quotes from funders

"What use is preservation when nobody is able to access the research data?"

Interview Spanish Research Council (CSIC),

26 April 2010

Quotes from funders

"The carrot and stick method might prove useful when it comes to encouraging researchers to take care of digital preservation. Release 80% of the research funding upfront, while waiting with the last 20% until research data and publication are secured for the future."

Interview Dutch Ministery of Education, Culture & Science (OCW), 8 Feb 2010

Quotes from funders

"Ultimately, researchers strive for immortality. Should it not be left up to them to achieve it?"

Interview Czech Republic Research Academy (ASCR), 7 April 2010

Bringing it all together…

10 insights & recommendations

Insight 1: policies

Insight	Existing policies are mostly focused on open access rather than on long-term preservation. Open access and long-term preservation are also often confused.
Solution	**Awareness:** preservation is important regardless of the access entitlements. **Rules & best practices:** show scientific advantages of LTP for researchers, efficiency gains for funders, etc.
Who	National governments, EU

Insight 2: policies

Insight	Various policies for (long-term) access to research output are developed at the same time. Duplicated efforts result in little efficiency.
Solution	**Rules & best practices:** an integrated and international approach is desired. Identify and analyse needs of stakeholders, focus on shared interests.
Who	National governments, EU, data managers

Insight 3: policies

Insight	No commonly accepted procedure for certification and accreditation of safe and reliable repositories exist.
Solution	**Rules & best practices:** start with a basic set of requirements for becoming a trusted digital repository. Support the development of an international standard and certification process.
Who	National governments, EU, data managers

Insight 4: organisational

Insight	No clarity exists about who should do what in preservation. In other words, who is responsible or takes responsibility.
Solution	**Rules & best practices:** business models should be developed. Stakeholders can achieve much more collaborative efforts (e.g. linking publications and data). Best practices can be taken from the publishers and libraries. Boundary conditions should be defined (legislation). **Funding:** a transition from project funding to structural funding for digital preservation should be made.
Who	EU, national governments, industry, funders

Insight 5: organisational

Insight	Some countries are far behind compared to others.
Solution	**Awareness:** make organisations aware that it is not only a technical challenge. **Rules & best practices:** best practices in preservation should be shared. A cross-domain virtual platform could help to learn about best practices in sharing and archiving of data.
Who	National governments, EU

Insight 6: organisational

Insight	Initiatives on preservation are taking place worldwide. Alignment is not yet reached.
Solution	**Awareness:** reach out and ensure cross-continental collaboration with projects and organisations such as in US, Asia and Australia. **Funding:** assign special budget for international cooperation (outside EU).
Who	EU, data managers, funders

Insight 7: infrastructure

Insight	Weak coverage of digital repositories. In some research disciplines no repositories are present at all.
Solution	**Awareness:** raise awareness of available repositories. **Rules & best practices:** agreements between research institutes and data managers for archiving research data. Develop common standards. **Technical infrastructure:** develop more trusted digital repositories that can deal with discipline specific data. **Funding:** inject money into trusted organisations on national level or domain specific for the creation of trusted repositories.
Who	EU, national governments, data managers, funders

Insight 8: infrastructure

Insight	Researchers are reluctant to share their data.
Solution	**Awareness:** encourage researchers to share their data, such as giving the possibility to receive credits. Else, use sticks(!) **Rules & best practices:** define and apply standards for openness and exchange of data (aka *Creative Commons* for data). Demonstrate linking and citability of data sets within and across disciplines. Implement DRM techniques to ensure that only those allowed can access the data. **Funding:** develop business models such as additional services.
Who	National governments, data managers, industry

Insight 9: infrastructure

Insight	Amount and variety of data will grow explosively. But not all data needs to be preserved. There is no consensus on what to keep for the long term.
Solution	**Rules & best practices:** commission expert panels (for each discipline) to support the selection process of what needs to be preserved and what not. Train researchers. Learn about collection build-up and management taking place at archives and libraries. Define the role of data curator. **Technical infrastructure:** ensure scalable solutions for storage, processing capacity and network bandwidth. **Funding:** funding should scale similar to the technical needs of the infrastructure.
Who	Nat. gov., EU, data managers, researchers

Insight 10: funding

Insight	Funders do not fully use their potential in shaping preservation activities.
Solution	**Rules & best practices:** funders should provide procedures to researchers on how preservation is to be given shape. Funders should incorporate preservation activities as part of project proposals. National governments and the EU should take on their responsibility and budget for preservation activities.
Who	Funders, national governments, EU

How we managed all data…

SurveyMonkey - Survey Results - Mozilla Firefox

Bestand Bewerken Beeld Geschiedenis Bladwijzers Extra Help

http://www.surveymonkey.com/MySurvey_Responses.aspx?sm=wN_2fPOc4uYfwgz5sK55KS

SurveyMonkey - Survey Results

SurveyMonkey

Signed in as "parsekb" | Sign Out

+ Create Survey | My Surveys | Address Book | My Account | Help

Digital research data survey (by PARSE.insight) Edit

Design Survey | Collect Responses | Analyze Results

- View Summary
- Browse Responses
- Filter Responses
- Crosstab Responses
- Download Responses
- Share Responses

Default Report + Add Report

Response Summary

Active Filter: Country filter

Total: 1,328 Edit Filter

Filtered: 48 Unapply Filter

Select a page to view below or view all pages:

« #8. Data re-use (cont.) »

PAGE: DATA RE-USE (CONT.)

14. Would you like to make use of research data gathered by other researchers in your discipline? Create Chart Download

	Response Percent	Response Count
Yes	60.0%	6
No	40.0%	4
answered question		10
skipped question		38

Swivel | PARSE.insight - Mozilla Firefox

Bestand Bewerken Beeld Geschiedenis Bladwijzers Extra Help

swivel.com https://www.swivel.com/people/1015959-PARSE-insight

PARSE.insight | Permanent Acce... Swivel | PARSE.insight

swivel

Home Reports Recent My Files Groups Public

Search Swivel

Feedback PARSE.insight Settings Sign out

SECTIONS

Overview

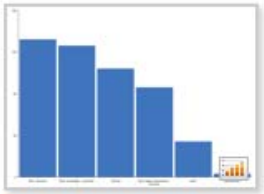
Activity

FILTERS

Public

PARSE.insight

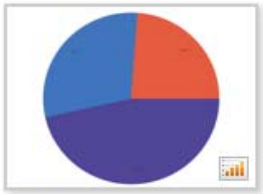
Public view all »



DATA MANAGERS - Apart from an

Posted by PARSE.insight

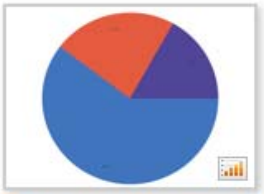
Modified Mar 30, 2010 9:48pm



DATA MANAGERS - Do you have other

Posted by PARSE.insight

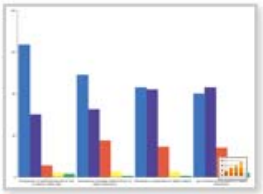
Modified Mar 30, 2010 9:48pm



DATA MANAGERS - Do you think that an

Posted by PARSE.insight

Modified Mar 30, 2010 9:48pm



DATA MANAGERS - Do you think the following

Posted by PARSE.insight

Modified Mar 30, 2010 9:48pm

Recent activity view all »

DATA MANAGERS - Apart from an infrastructure, what do you...
Updated 3 months ago

DATA MANAGERS - Do you have other arrangements/means for ...
Updated 3 months ago

DATA MANAGERS - Do you think that an international infras...
Updated 3 months ago

DATA MANAGERS - Do you think the following initiatives wo...
Updated 3 months ago

DATA MANAGERS - How Important do you regard the following...
Updated 3 months ago

DATA MANAGERS - Please indicate how important you think t...
Updated 3 months ago

DATA MANAGERS - Regarding current movements in scholarly ...
Updated 3 months ago

DATA MANAGERS - In your opinion, who is responsible for p...
Updated 3 months ago

Questions? Suggestions?

Send us feedback

Send

Contact us

@teamswivel on Twitter

support@swivel.com

Learn more

Tour

Features

Pricing

About

Company

Jobs

Blog

Help

FAQ

Developers

Klaar

Interactive Map of actors and stakeholders in Digital Preservation - Mozilla Firefox

Bestand Bewerken Beeld Geschiedenis Bladwijzers Extra Help

http://www.parse-insight.eu/imap/imap.html

Interactive Map of actors and ...

Interactive Map of actors and stakeholders in Digital Preservation

Show Organisations | Help

Select all | Deselect all

- Funding/policy
- Research
- Preservation
- Publishing

Please select the **categories** to show on the map. Click on the markers on the map, or hover your mouse to get a tooltip with a brief description.

Select All Europe | Deselect all

- Korea
- Philippines
- Taiwan
- Uruguay
- Liechtenstein

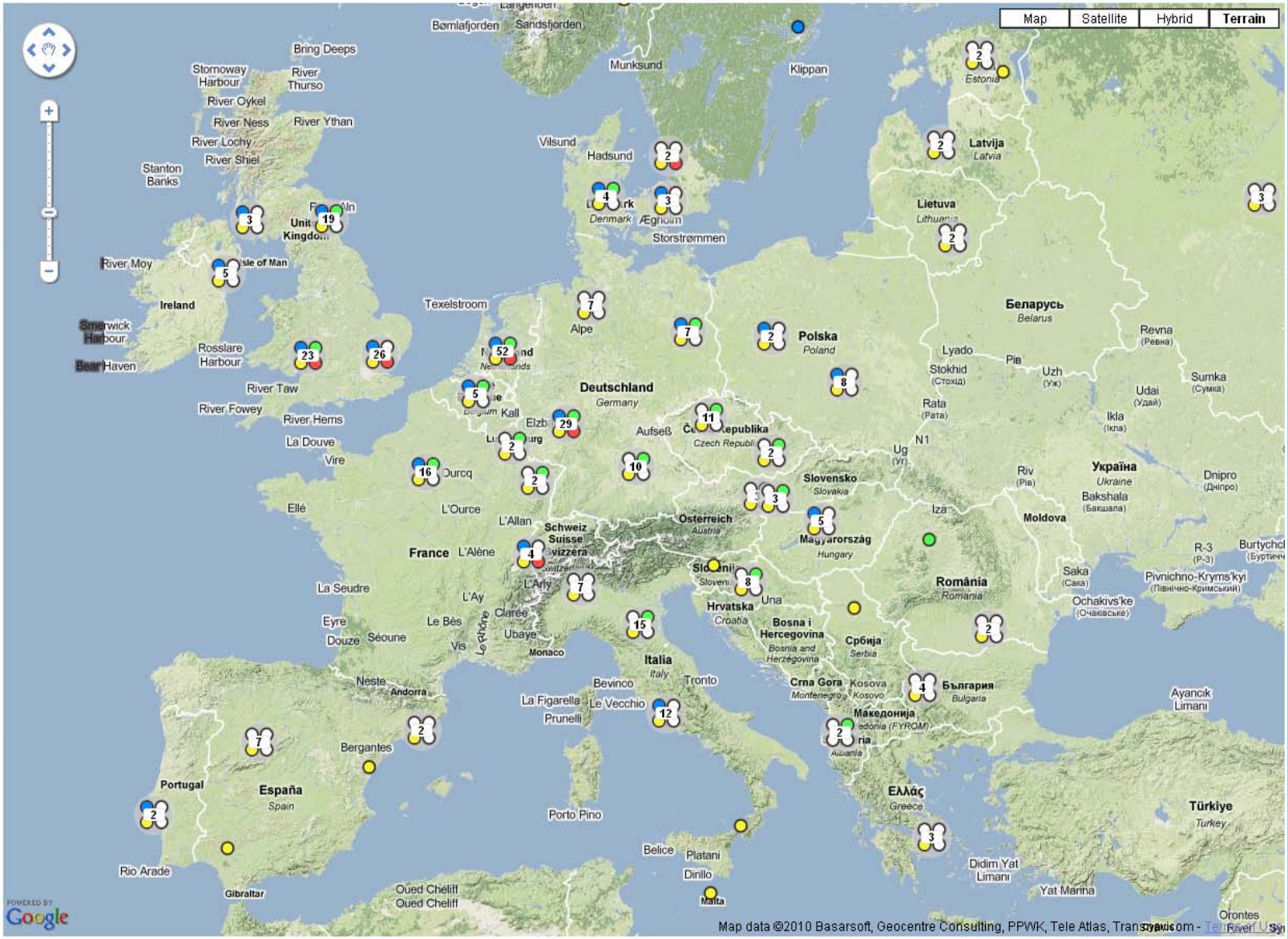
Only organisations from selected **countries** will be shown on the map.

Select all | Deselect all

- Agriculture and nutrition
- Behavioral sciences
- Humanities
- Life sciences
- Medicine
- Physical sciences
- Socio-cultural sciences
- Social sciences
- Technology
- Psycholinguistic
- <no Community>

Please select the **communities** to show on the map.

info@parse-insight.eu



Klaar

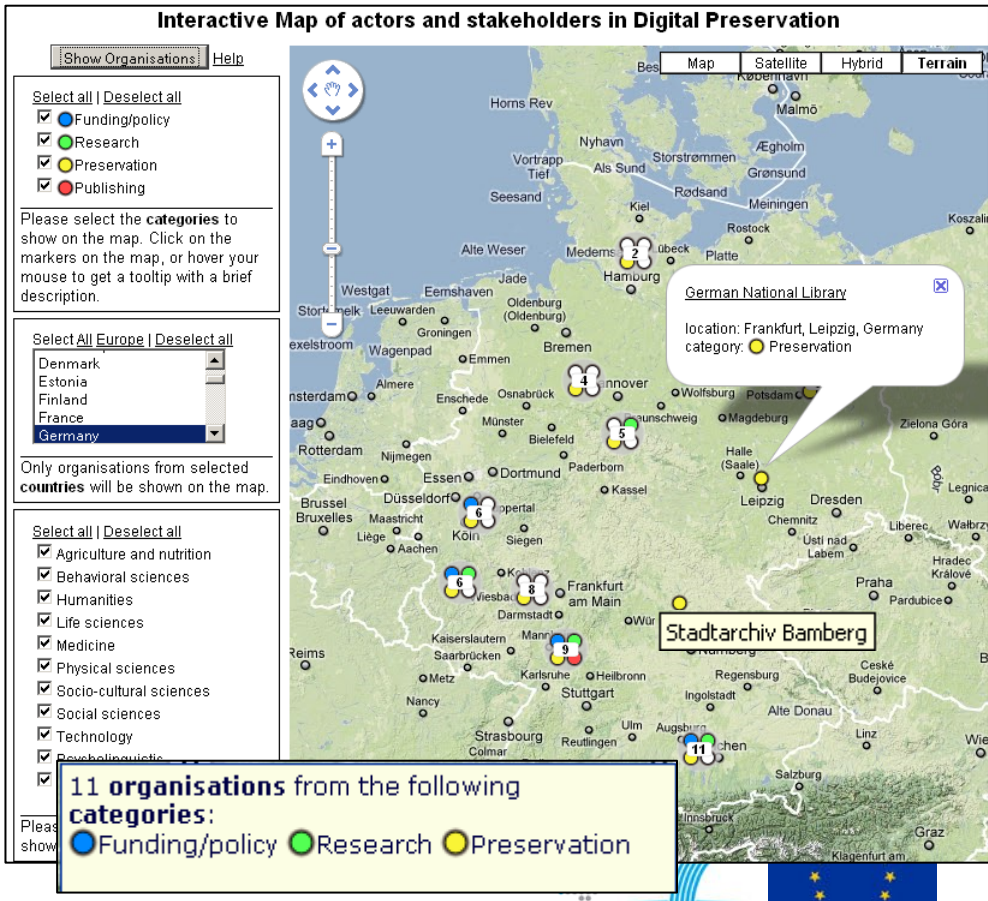
Interactive Map

Conceptual Outline and User Interface

- Webpage to locate ‘actors’ on map
- Aggregate icons provide overview:
  - Number of clustered actors
  - Contained categories
  - Tooltip for explanation
- Drill-Down using aggregate icons
- Color coded icons:

Funding/policy
Research
Preservation
Publishing

- Bubbles provide detailed info per actor
- Tooltip displays name
- Simple but powerful search feature



Most important insight reports

- D3.3 Case studies report
- D3.4 Survey report
- D3.5 Interim Insight report
- D3.6 Final Insight report

These reports are available on our website:

www.parse-insight.eu

Thank you!

e-infrastructure

SEVENTH FRAMEWORK PROGRAMME