Project

Lourdes Verdes-Montenegro & AMIGA team

Instituto de Astrofísica de Andalucía (CSIC)

19th October 2017, AENEAS all-hands meeting (IAA, Granada)

SKA-LINK Project

...IS NOT THE FUNDING FOR SPAIN TO JOIN SKA!!!

(sorry)

Lourdes Verdes-Montenegro & AMIGA team

Instituto de Astrofísica de Andalucía (CSIC)

19th October 2017, AENEAS all-hands meeting (IAA, Granada)

Project

Is about science that can be trusted in N yrs

Lourdes Verdes-Montenegro & AMIGA team

Instituto de Astrofísica de Andalucía (CSIC)

19th October 2017, AENEAS all-hands meeting (IAA, Granada)

- Funded by the i-LINK CSIC program for international scientific collaborations (participants list later)
- 2017 & 2018
- Overall aim:

Contribute to make SKA a **reference** not only in science and technology but in **scientific methodology**, by

producing a general framework of Best Practices to be considered in the design of the SKA Regional Centres

SRCs as Open Science Hubs

- SRCs will constitute a service to the community:
  - "...users will have access to data products they are authorised to, as well as the tools and processing power to generate and analyse advanced data products"

SRCCG Document *SKA Regional Centres: Background And Framework*

Platforms to share data, methods and knowledge --- Open Science Hubs

SRCs as Open Science Hubs

- SRCs will constitute a service to the community:
  - "...users will have access to data products they are authorised to, as well as the tools and processing power to generate and analyse advanced data products"

SRCCG Document *SKA Regional Centres: Background And Framework*

Platforms to share data, methods and knowledge --- Open Science Hubs

Extract scientific knowledge from such data deluge:

"*If there is a data deluge then there is also a deluge in the methods used to process it*"

De Roure & Goble 2010, Anchors in Shifting Sand: the Primacy of Method in the Web of Data

SRCs as Open Science Hubs

- SRCs will constitute a service to the community:
  - "...users will have access to data products they are authorised to, as well as the tools and processing power to generate and analyse advanced data products"

SRCCG Document *SKA Regional Centres: Background And Framework*

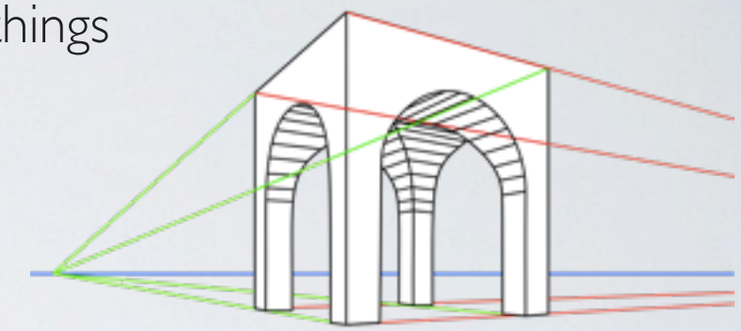
Platforms to share data, methods and knowledge --- Open Science Hubs

Extract scientific knowledge from such data deluge:

"*If there is a data deluge then there is also a deluge in the methods used to process it*"

De Roure & Goble 2010, Anchors in Shifting Sand: the Primacy of Method in the Web of Data

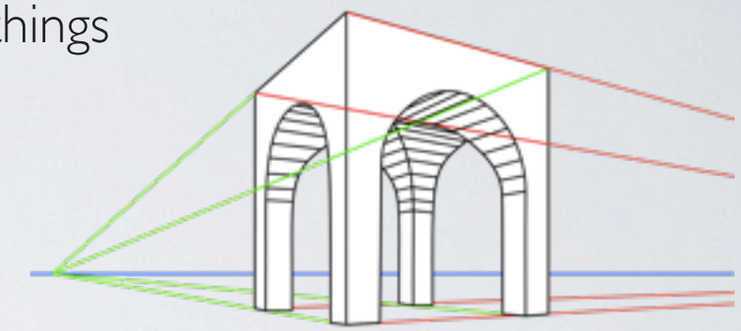
PERSPECTIVES



Data to the desktop (individual users perspective)

- because I have the best code, which I know how to use and can do special things
- because I do not trust any "pipeline" that you made
  - partly because I know better how to do it
  - partly because I read the news and there is a reproducibility crisis
  - well, and myself I hardly can reproduce the results of other's papers...
- in general I want full control of the software and of the computational environment

PERSPECTIVES



Data to the desktop (individual users perspective)

- because I have the best code, which I know how to use and can do special things
- because I do not trust any "pipeline" that you made
  - partly because I know better how to do it
  - partly because I read the news and there is a reproducibility crisis
  - well, and myself I hardly can reproduce the results of other's papers...
- in general I want full control of the software and of the computational environment

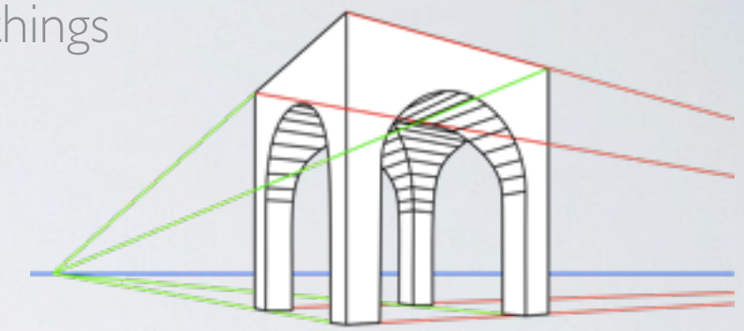
Computation to data - to fully exploit SKA (providers perspective)

- we need to install your software in the SRC platform, can we trust it?, can we run it?, environment, dependencies

- Hey, we are offering services to the community, computation + tools. We would be grateful if you allow us to share it with other users (with proper credit)

- Mmmm, sharing is great, but, putting the software in the platform is not enough: you need to provide the context for people to be able to rerun it in the same or other data

PERSPECTIVES



Data to the desktop (individual users perspective)

- because I have the best code, which I know how to use and can do special things
- because I do not trust any pipeline that you made
  - partly because I know better how to do it
  - partly because I read the news and there is a reproducibility crisis
  - well, and myself I hardly can reproduce the results of other's papers...
- in general I want full control of the software and of the computational environment

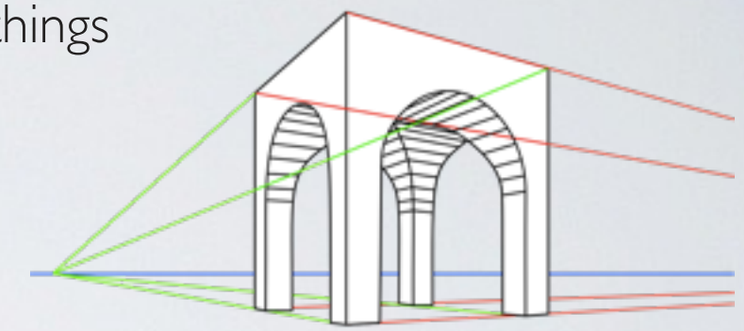
Consortia of users -- "KSP teams"

- we have tools to generate ADPs, and we will put them in the SRCs because there is where the storage and computation is
- but... we put effort on it, what would I gain if I make the *additional effort* to make it reusable?
- well, maybe we will be share in 4 yrs time (PhD typical time)
- If I make it, then I will pave the way to competitors
- Pressure to "make the discovery": Publish or perish

Computation to data - to fully exploit SKA (providers perspective)

- we need to install your software in the SRC platform, can we trust it?, can we run it?, environment, dependencies
- Hey, we are offering services to the community, computation + tools. We would be grateful if you allow us to share it with other users (with proper credit)
- Mmmm, sharing is great, but, putting the software in the platform is not enough: you need to provide the context for people to be able to rerun it in the same or other data

PERSPECTIVES



Data to the desktop (individual users perspective)

- because I have the best code, which I know how to use [illegible] things
- because I do not trust any pipeline that you
  - partly because I know better how to do it
  - partly because I read the news and there is a reproducibility crisis
  - well, and myself I hardly can reproduce the results of other's papers...
- in general I want full control of the software and of the computational environment

This is about reuse and trust

Consortia of users -- "KSP teams"

- we have tools to generate ADPs, and we will put them in the SRCs [illegible] storage and computation is
- but... we put effort on it, what would I gain if I make the *additional ef[illegible] to make it reusable?
- well, maybe we will be share in 4 yrs time (PhD typical time)
- If I make it, then I will pave the way to competitors
- Pressure to "make the discovery": Publish or perish

This is about metrics of science

Computation to data - to fully exploit SKA (providers perspective)

- we need to install your software in the SRC platform, can we trust it, can we [illegible] dependencies
- Hey, we are offering services to the community, computation + tools. We wou[illegible]llow us to share it with other users (with proper credit)
- Mmmm, sharing is great, but, putting the software in the platform is not enough: you need to provide the context for people to be able to rerun it in the same or other data

This is about technology (and AENEAS)

SKA-LINK IN A NUTSHELL

- **Areas of work**

  - Facilitate the **reproducibility** of the scientific methods and their **verification**, then their **reuse** and **repurpose follows**. FAIR principles

  - Identification of **barriers** and ways to overcome them

  - Inventory of technologies/ technical strategies

  - Incentives/Metrics

- **How: collaboration among**

  - Members of the Science Data Processor (SDP) consortium

  - Experts involved in the design of the SKA Regional Centres

  - Specialists on e-Science technologies for the scientific exploitation of DCIs

- SDP / AENEAS members:
  - (PI) Lourdes Verdes-Montenegro (SRCCG), Julián Garrido, Susana Sánchez (IAA-CSIC)
  - Paul Alexander, Rosie Bolton (SRCCG), Bojan Nikolic (U. of Cambridge)
  - Anna Scaife, Chris Skipper (U. of Manchester)
  - Robert Simmonds, David Aikema, Adrianna Pinska (U. of Cape Town)
  - Andreas Wicenec (ICRAR)
  - Michael Wise (SRCCG), Yan Grange, Hanno Holties, Rob Van der Meer (ASTRON)
  - +(invited) Antonio Chrysostomou (SKAO/SRCCG), Russ Taylor, Severin Gaudet (SRCCG)

Contributions to SKA-Link are fully open

- European groups developing leading-edge e-Science technologies:
  - **Malcolm Atkinson, Rosa Filgueira, Amrey Krause (U. of Edinburgh)**. Major contributions to EU projects ADMIRE, ENVRI, EUDAT and VERCE, and it leads the design of the e-Infrastructure in the H2020 projects EUDAT20
  - **Peter Kacsuk, Zoltan Farkas (SZTAKI)**. Expert in developing generic science gateway frameworks based on workflows -gUSE/WS-PGRADE
- **Representatives from other communities that have applied e-Science technologies**
  - **Jens Krüger (U. of Tübingen).** Developer of the Science Gateway for the Molecular Biology community (MoSGrid)
  - **Alessandro Spinuso, Wim Som de Cerff (KNMI).** Leader of the design and the implementation of the VERCE Science Gateway for the Earth Science community.
  - **Rafael Garrido (IAA-CSIC).** Pioneer team in developing asteroseismic tools in the VO, ported as well to the Grid environment,

SKA-LINK IN A NUTSHELL

- **Areas of work**
  - Facilitate the **reproducibility** of the scientific methods and their **verification**, then their **reuse** and **repurpose follows**
  - Identification of **barriers** and ways to overcome them
  - Inventory of technologies/ technical strategies
  - Incentives/Metrics

- **How: collaboration among**
  - Members of the Science Data Processor (SDP) consortium
  - Experts involved in the design of the SKA Regional Centres
  - Specialists on e-Science technologies for the scientific exploitation of DCIs
- **SKA-link deliverables will be integrated into SRCCG milestones**
  - Set of best practices for SRCs to be considered a reference in scientific methodology
  - Potential tools to achieve it
  - Requirements / Goals

Any problem with the Scientific Method??

- Reuse requires reproducible, which, BTW is a principle of the Scientific Method (1660s)

ANY PROBLEM WITH THE SCIENTIFIC METHOD??

- Reuse requires reproducible, which, BTW is a principle of the Scientific Method (1660s)

25 May 2016

- Questionnaire on reproducibility filled by 1500 scientists
- > 70% of researchers have tried and failed to reproduce another scientist's experiments
- **> 50% have failed to reproduce their own experiments**
  - Chemistry: 90% (60%)
  - Biology: 80% (60%)
  - Physics and engineering: 70% (50%)
  - Medicine: 70% (60%)
  - Earth and environment science: 60% (40%)

ANY PROBLEM WITH THE SCIENTIFIC METHOD??

- Reuse requires reproducible, which, BTW is a principle of the Scientific Method (1660s)

25 May 2016

- Questionnaire on reproducibility filled by 1500 scientists
- > 70% of researchers have tried and failed to reproduce another scientist's experiments
- **> 50% have failed to reproduce their own experiments**
  - Chemistry: 90% (60%)
  - Biology: 80% (60%)
  - Physics and engineering: 70% (50%)
  - Medicine: 70% (60%)
  - Earth and environment science: 60% (40%)

Ah! So you don't empathise?

ANY PROBLEM WITH THE SCIENTIFIC METHOD??

- Reuse requires reproducible, which, BTW is a principle of the Scientific Method (1660s)

25 May 2016

- Questionnaire on reproducibility filled by 1500 scientists
- > 70% of researchers have tried and failed to reproduce another scientist's experiments
- **> 50% have failed to reproduce their own experiments**
  - Chemistry: 90% (60%)
  - Biology: 80% (60%)
  - Physics and engineering: 70% (50%)
  - Medicine: 70% (60%)
  - Earth and environment science: 60% (40%)

Overly Honest Method
@OverlyHonestly

Maybe with this?

You can download our code from the URL supplied. Good luck downloading the only postdoc that can get it to run, though #OverlyHonestMethods

SKA-LINK PROGRESS

Working definitions, within the context of SRCs

- **Reproducibility (in theory):** An experiment/study is reproducible if an external researcher could repeat the same procedures and confirm the results using the same set up, input data and methods.

SKA-LINK PROGRESS

Working definitions, within the context of SRCs

- **Reproducibility (in theory):** An experiment/study is reproducible if an external researcher could repeat the same procedures and confirm the results using the same set up, input data and methods.

- **Reproducibility (in practice):** input data, methods, set up parameters, output data and results, together with details on the context and links between the pieces of the experiment.

  - researchers can discover and manage (big) scientific data. This requires: standards for data interoperability, metadata for the data collections, storage infrastructures, annotations.

  - researchers can replay the software applications. ...measures to avoid software decay, access to DCIs, etc.

  - researchers can validate and trace their experiments. Provenance-aware technology.

  - researchers can collaborate and share their experiments. Collaborative framework that facilitates the acknowledgement of work.

SKA-LINK PROGRESS

- **Scientific methods vs Software:** software together with the corresponding context that describes the inputs, outputs as well as the analytical tasks implemented by this software.

  It depends on the annotation capacity of the platform

- **Scientific workflows vs Pipelines:**

  - pipelines are more focused on producing scientifically exploitable data products (end-to-end) and are often black boxes

  - scientific workflows expose the structure of the software, inputs and outputs

Inventory of technologies

- **Science Gateways**
    - Technologies: WS-PGRADE/gUSE, hubzero, iRODS.
    - Examples/Use cases: VERCE, MoSGrid, CyberSKA, VIALACTEA.
- **Research repositories for scripts, code workflows, ROs and other research pieces**: GitHub, MyExperiment, ROHub, ASCL, Zenodo
- **Workflows management:** dispel4py, Asterism, Taverna, Pegasus
- **Virtualization:** Singularity, rocket, runC and Docker containers, Skyport, Occopus, Virtual Machines, Sandboxing
- **Interactive computing:** Jupyter notebooks
- **Data sharing**: data discovery and access services, implemented based on IVOA standards. ONEDATA. Data models
- **Technologies for modelling / capturing / visualizing experiments and their provenance:**
    - Ontologies, e.g. RO model, PROV data model, S-PROV model
    - Tools, e.g. S-ProvFlow, provenance store and conversion/validation services.
- **IaaS tools**, deploy and manage virtual infrastructures (example workflows) in clouds in a portable way). OpenStack

Barriers: experience from other communities

1. What needs led this community to adopt the Science Gateways technology?
    - way for a transparent representation of the recipes
    - a platform to perform the calculations and to store the data
    - mutual exchange
    - improved transparency and trust within a community.
    - Abstract the complexity of clouds and HPC
2. Process leading to adoption by the community?
    - power users sharing their workflows with the purpose of training
    - evaluation of the status of different gateways by a technical team
    - ....

Barriers: experience from other communities

3. Barriers found during the process of adoption
    - Complexity of setting up a production deployment
    - Updating the science gateway as new releases came up
    - Learning curve and usage
    - Authentication procedures are a major obstacle
    - Propagation of error message through the complex software stack of science gateway
    - Granularity and usability of the provenance
4. How were these barriers lowered?
    - Reactiveness of the gUSE team to support request and bug fixes
    - Manual upgrading process
    - Training and documentation
    - Direct communication with the experts of the gateway technology and with other communities
    - Single-sign-on has become a de facto standard
    - Solutions such as semantic workflow description
    - Support from the technology providers

AND ALL THIS COMES BACK TO METRICS

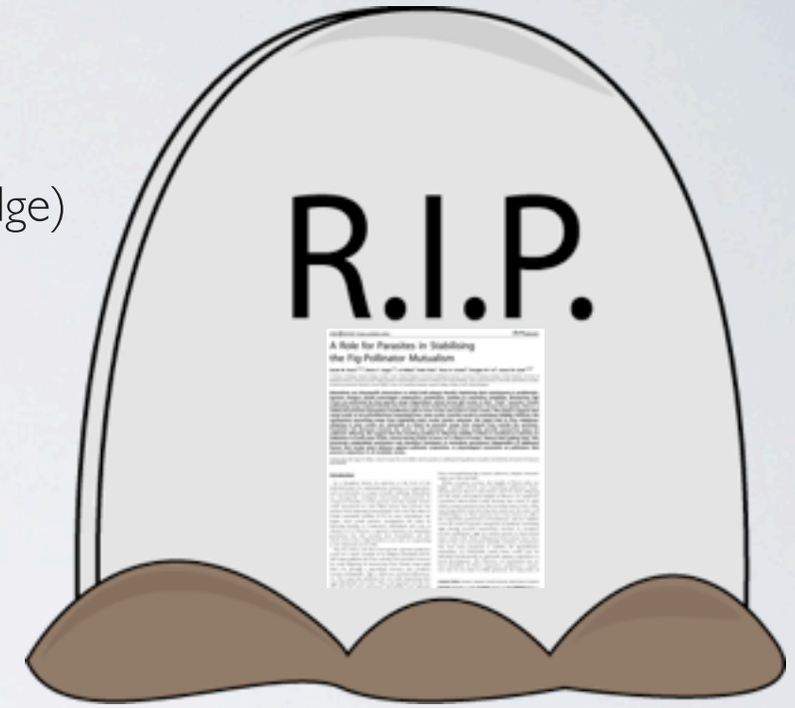
- **Knowledge Burying in paper publication**

(S. Bechhofer 2011, Research Objects: Towards Exchange and Reuse of Digital Knowledge)

- Publishing/mining cycle results in loss of knowledge

>= 40% of information lost

- **RIP**: **R**est **I**n **P**aper



http://www.clipartkid.com/rip-cliparts/

Moving from narratives (last 300 yrs)
to the actual output of research

METRICS

nature International weekly journal of science

SPECIALS

See all spec

How to improve the use of metrics

Nature **465**, 870–872 (17 June 2010) | doi:10.1038/465870a

... "Science is being killed by numerical ranking,"[...] Ranking systems lures scientists into pursuing high rankings first and good science second.

SCIENCE METRICS

The value of scientific output is often measured, to rank one nation against another, allocate funds between universities, or even grant or deny tenure. Scientometricians have devised a multitude of 'metrics' to help in these rankings. Do they work? Are they fair? Are they over-used? *Nature* investigates.

▼ Editorial ▼ Features ▼ Opinion ▼ From the archive

EDITORIAL

Assessing assessment

Transparency, education and communication are key to ensuring that appropriate metrics are used to measure individual scientific achievement.

METRICS

nature International weekly journal of science

SPECIALS

See all spec

How to improve the use of metrics

Nature **465**, 870–872 (17 June 2010) | doi:10.1038/465870a

... "Science is being killed by numerical ranking,"[...] Ranking systems lures scientists into pursuing high rankings first and good science second.

SCIENCE METRICS

The value of scientific output is often measured, to rank one nation against between universities, or even grant or deny tenure. Scientometricians have c 'metrics' to help in these rankings. Do they work? Are they fair? Are they over- investigates.

▼ Editorial ▼ Features ▼ Opinion ▼ From the archive

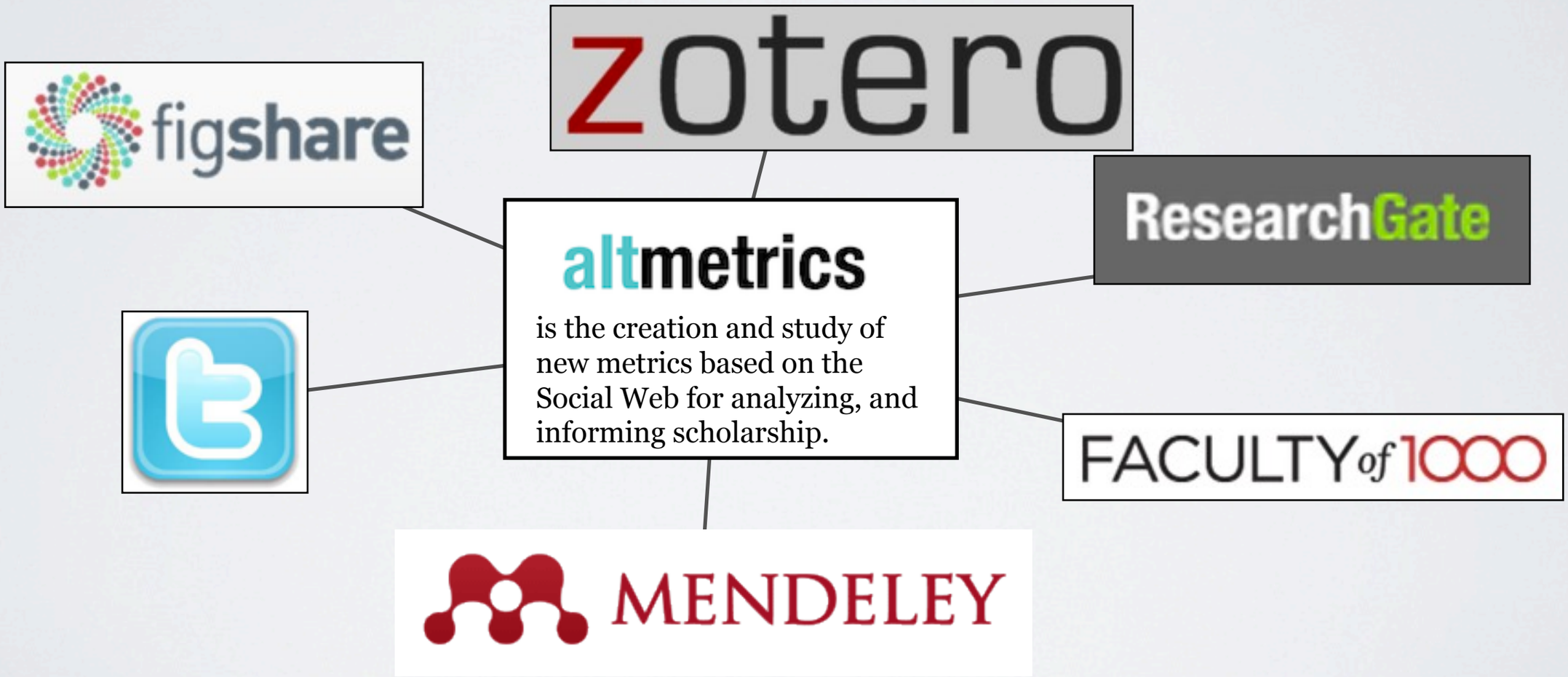
EDITORIAL

Assessing assessment

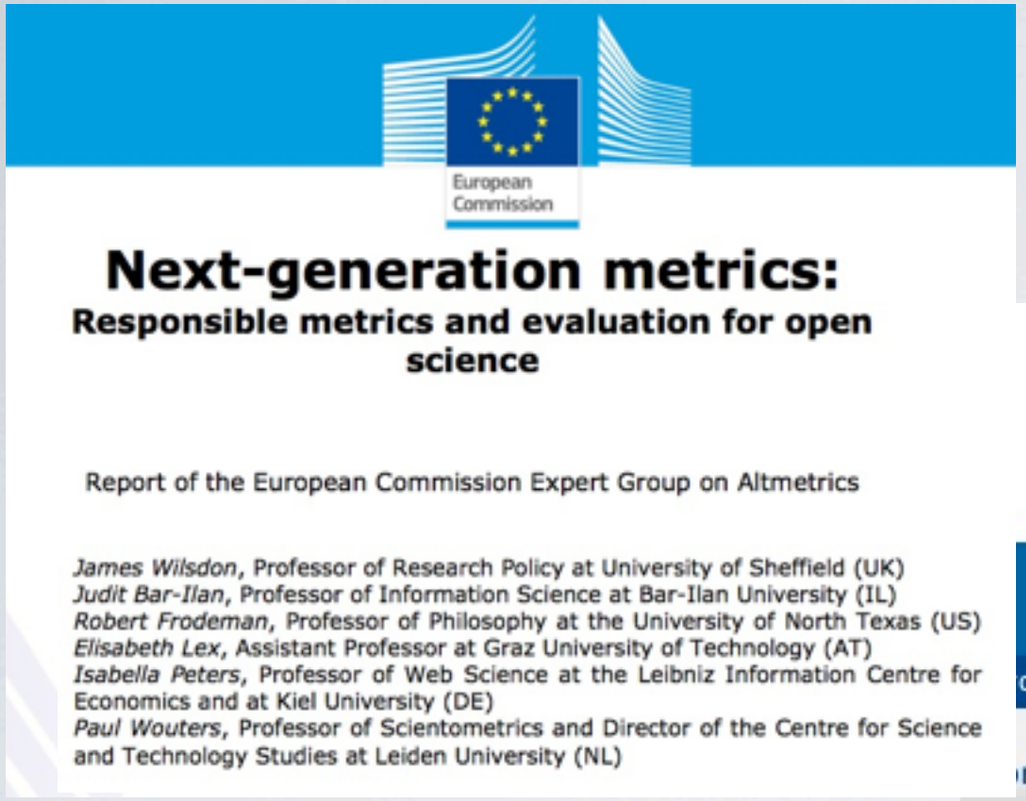
Transparency, education and communication are key to ensuring that appropriate metrics are used to measure individual scientific achievement.

METRICS

- Citations represent less than 1% of usage for an article



METRICS



European Commission

Next-generation metrics:
Responsible metrics and evaluation for open science

Report of the European Commission Expert Group on Altmetrics

James Wilsdon, Professor of Research Policy at University of Sheffield (UK)
Judit Bar-Ilan, Professor of Information Science at Bar-Ilan University (IL)
Robert Frodeman, Professor of Philosophy at the University of North Texas (US)
Elisabeth Lex, Assistant Professor at Graz University of Technology (AT)
Isabella Peters, Professor of Web Science at the Leibniz Information Centre for Economics and at Kiel University (DE)
Paul Wouters, Professor of Scientometrics and Director of the Centre for Science and Technology Studies at Leiden University (NL)

European Commission

RESEARCH & INNOVATION

Open Science

Expert Group on Altmetrics

NEW: Final Report of the Expert Group on Altmetrics is available

Publication date: 20 March 2017

The Expert Group on Altmetrics outlines in this report how to advance a next-generation metrics in the context of Open Science and delivers an advice corresponding to the following policy lines of the Open Science Agenda: Fostering Open Science, Removing barriers to Open Science, Developing research infrastructures and Embed Open Science in society.

The report will be presented and discussed at the Open Science Policy Platform on 20 March 2017

Altmetrics can stimulate the adoption of open science principles, i.e., collaboration, sharing, networking. Altmetrics also have potential in the assessment of interdisciplinary research and the impact of scientific results on the society as a whole, as they include the views of all stakeholders and not only other scholars (as with citations).

METRICS

Euroscience Open Forum (ESOF) 2018:

- Session proposed for "Theme #3 Science policy and transformation of research practice"

  "*Is the current measure of excellence perverting Science? A Data deluge is coming, it is time to act*"

- Focused on reproducible science and new metrics in the era of Megascience infrastructures
- Current metrics to measure success of contracts, grants, teams, institutes or scientific infrastructures do not help: based on number of papers/citations (< 1% of usage for an article) they do not promote reproducibility, so that excellence in science is being killed by numerical ranking

METRICS

Euroscience Open Forum (ESOF) 2018:

- Participants
  - William Garnier (SKAO) - Submitter and Manager
  - May Chiao (Chief Editor Nature Astronomy) - Moderator
  - Keynote speakers
    - Sebastian Neubert (Univ. of Heidelberg, involved in the Worldwide LHC Supercomputing Grid
    - Lourdes Verdes-Montenegro (IAA-CSIC)
- Panellists:
  - Carole Goble (Univ. of Manchester, working with many different scientific communities towards reproducible research)
  - Jeff Dozier (Univ. of California, applies reproducibility in climate change studies)
  - René Von Schomberg (Leads the EC Open Science policy coordination and development team)
  - Antonio Chrysostomou (SKAO)

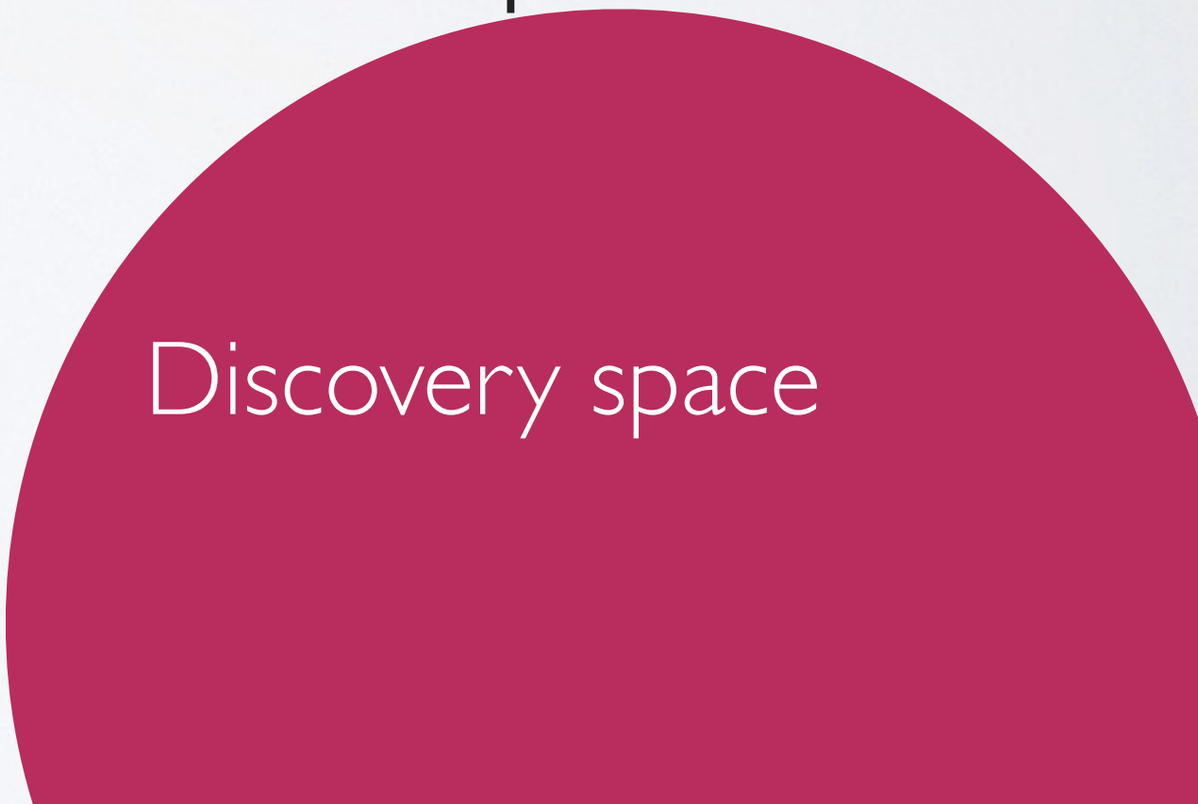
BENEFITS FOR THE SKA COMMUNITY

- Reproduciblity is not the aim, is the mean

- SRCs synonymous of Science as a Service (SClaaS)? (not meaning outsourcing)

  - Supporting scientific communities to access, share, and reuse research objects, methods, experiments, stimulating the development of new knowledge

- Keeping a project at the scale of the SKA funded requires all of the science to be spotless



Rosie Bolton slides

An opportunity for SKA?