

EDISON: Coordination and cooperation to establish new profession of Data Scientist for European Research and Industry

Yuri Demchenko
University of Amsterdam



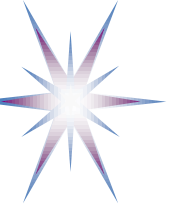
EDISON
building the data
science profession

Data and computing infrastructures for open
scholarship Workshop

22 September 2015, RDA6, Paris

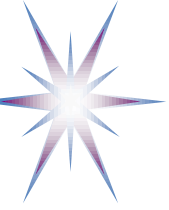
EDISON – **E**ducation for **D**ata Intensive
Science to **O**pen **N**ew science frontiers

Grant 675419 (INFRASUPP-4-2015: CSA)



Outline

- Consortium members
- EDISON Project Objectives
- Project structure – WPs, Tasks and main products
- EDISON Approach:
 - eCFv3.0 and CF-DS definition
 - DS-BoK and Model Curriculum
 - Education and Training Infrastructure for Data Science
- EDISON and e-Infrastructure for Open Scholarship
 - Today's questions - Topics for discussion



EDISON Consortium Members

Universities

1. **University of Amsterdam (UvA) - Coordinator**
2. University of Stavanger (UiS)
3. University of Southampton (UK)

Community related partners

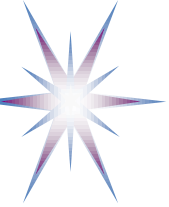
4. European Grid Initiative (EGI)
5. FTK – Research Institute for Telecommunication and Cooperation (DE) and APARSEN (Alliance Permanent Access to Records of Science in European Network)

Industry partner

6. Industry: Engineering (Italy)

SME

7. InMark (Spain)



Project Objectives

Objective 1: Promote the creation of Data Scientists curricula by an increasing number of universities and professional training organisations.

Define common framework for building new curricula on Data Science including Data Science Competence Framework (CF-DS), Body of Knowledge (DS-BoK) and Model Curriculum (MC-DS).

Work with champion universities and organisations to implement MC-DS

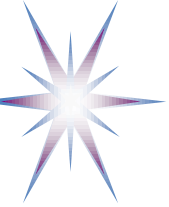
Objective 2: Provide conditions and environment for re-skilling and certifying Data Scientists expertise to graduates, practitioners and researchers throughout their careers.

Support Data Science education and training for students and “self-made” Data Science practitioners to allow their formal professional certification.

Create EDISON Education and Training Marketplace by leveraging EGI Engage Training Marketplace.

Objective 3: Develop a sustainable business model and a roadmap to achieve a high degree of competitiveness for European education and training on Data Science technologies, provide a basis for the formal recognition of the Data Scientist as a new profession.

Create Community of practice for sustainable Data Science education and training supported by EDISON Liaison Group(s).

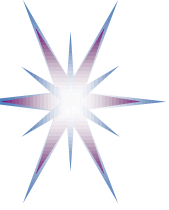


Objectives 1: Data Science Profession Definition: Competences Framework and Model Curriculum

Objective 1: Promote the creation of Data Scientists curricula by an increasing number of universities and professional training organisations.

EDISON contribution

- The ***Data Science Competence Framework (CF-DS)*** including Glossary and Taxonomy of competences and skills.
- The ***Body of Knowledge (BoK) for the Data Scientist Professional (DS-BoK)*** that will be used to map required competencies/skills and existing academic, research and technology disciplines
- A ***Model Curriculum for Data Science (MC-DS)*** that will provide a reference implementation of the proposed CF-DS and DS-BoK.

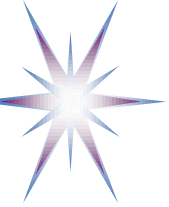


Objective 2: Education and Training Environment

Objective 2: Provide conditions and environment for re-skilling and certifying Data Scientists expertise to graduates, practitioners and researchers throughout their careers.

Provide conditions for a sustainable increase in the number of Data Scientists in Europe to satisfy the level of research and industry demand in coming 5-8 years.

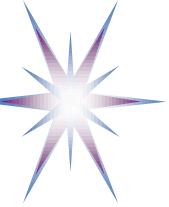
- ***Education and Training e-Infrastructures*** to support specialist training on Data Science that will include access to data sets for educational purposes, data management and analytical tools and cloud based virtual labs and classrooms.
- ***EDISON Marketplace*** of existing and developing education and training courses, education materials and other resources, by leveraging on the EGI Knowledge Commons and the Training Market initiative.
- A model and a framework for ***Data Science education programme accreditation and professional certification***.



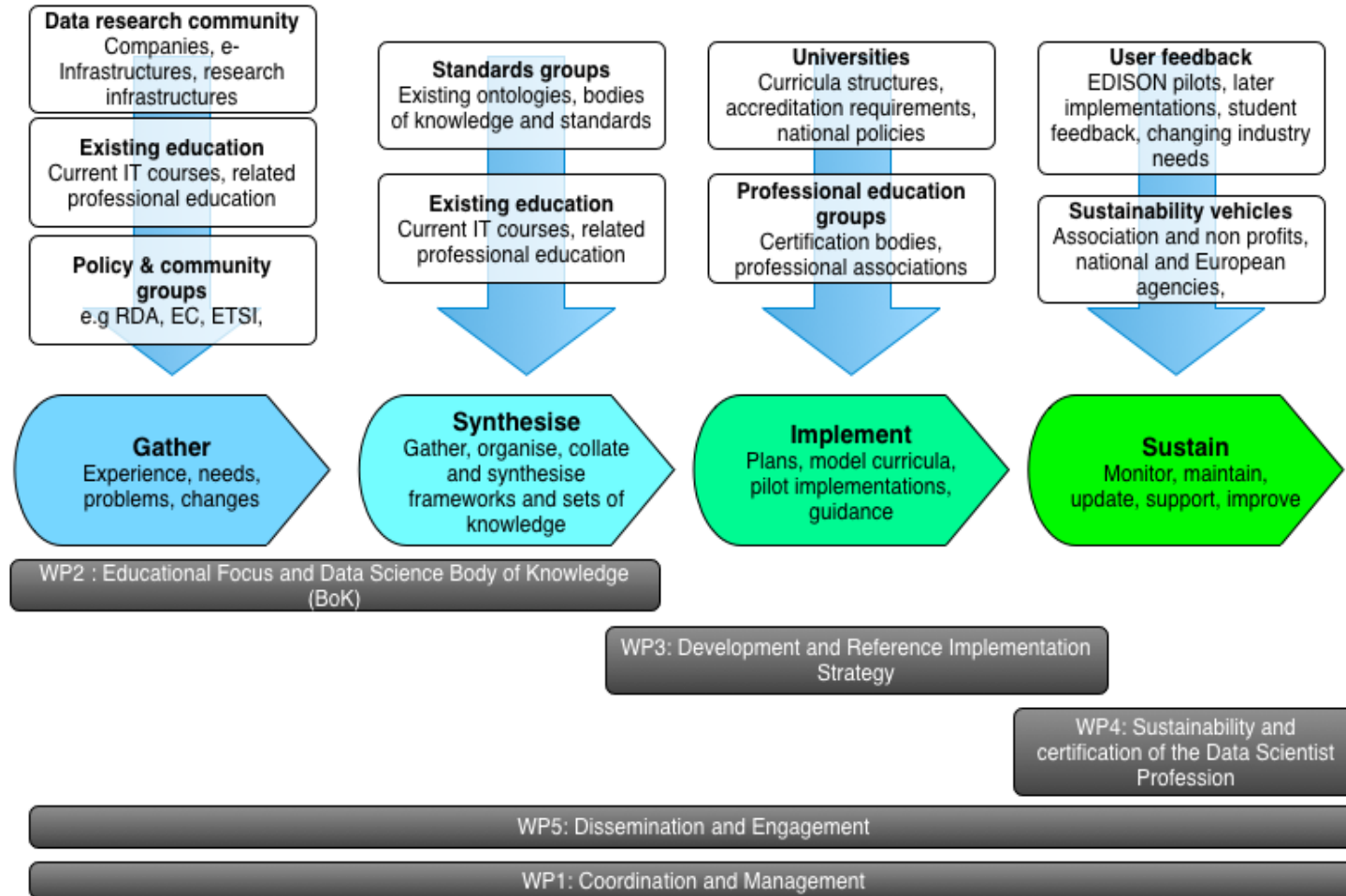
Objective 3: Sustainability model

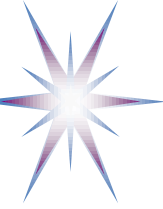
Objective 3: Develop a sustainable business model and a roadmap to achieve a high degree of competitiveness for European education and training on Data Science technologies, provide a basis for the formal recognition of the Data Scientist as a new profession.

- An ***education and training model*** that will include multiple learning models (e.g. residential, lifelong learning, e-Learning) that will allow ***adjusting individual career path*** in Data Science.
- A ***business model*** that incorporates the proposed education model into a sustainable cycle of the research data value chain including the major stakeholders and actors from research, industry and government.
- A ***roadmap for a sustainable education and training services*** as a part of overall data driven technologies development in Europe.
- ***EDISON Liaison Group(s) (ELG)*** of independent experts that represent the major stakeholders in Data Science that will work as a consulting body for the project and will create a basis for *future independent expert group for universities and for EC.*

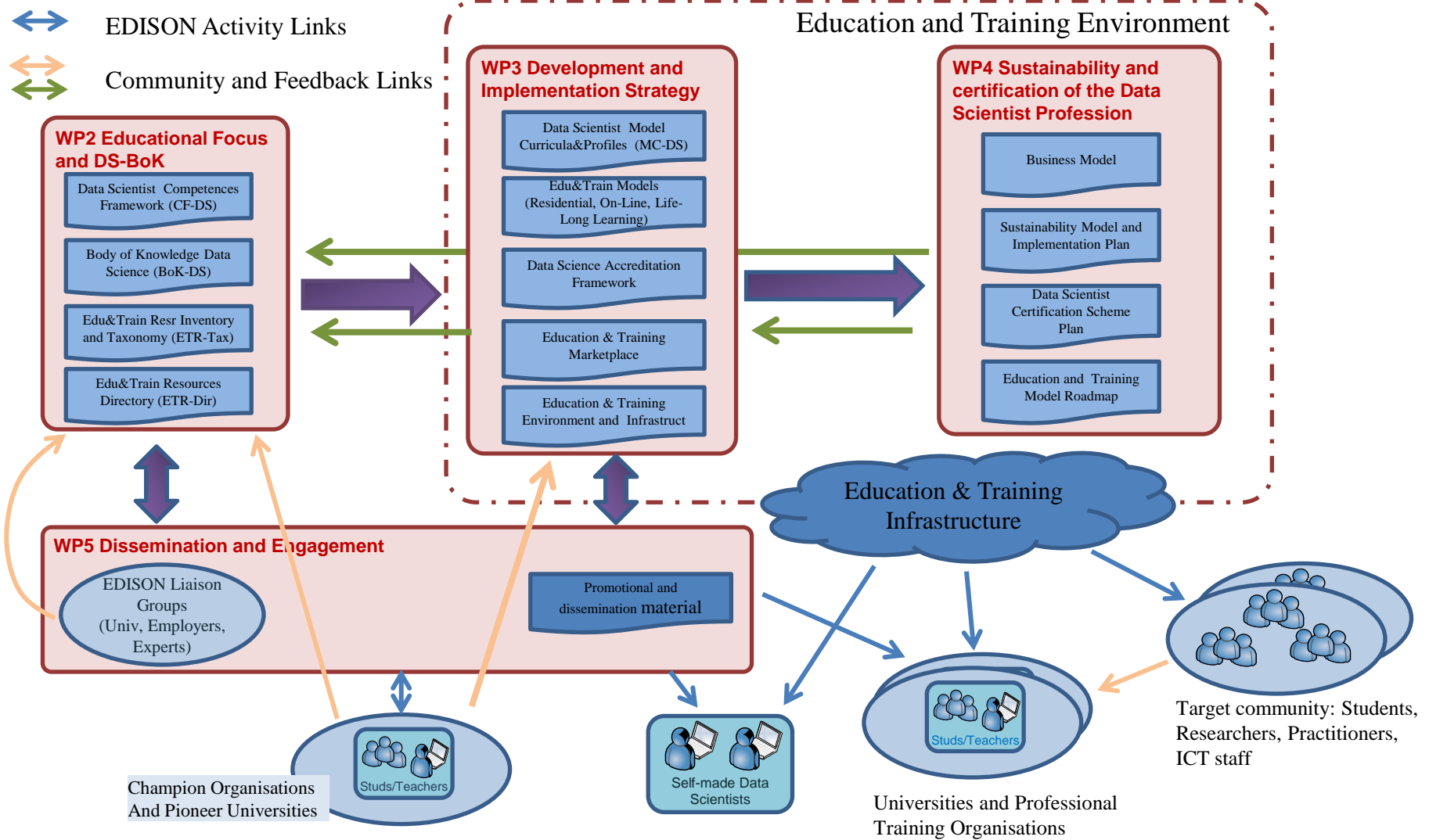


Basic methodology of EDISON, inputs and work packages



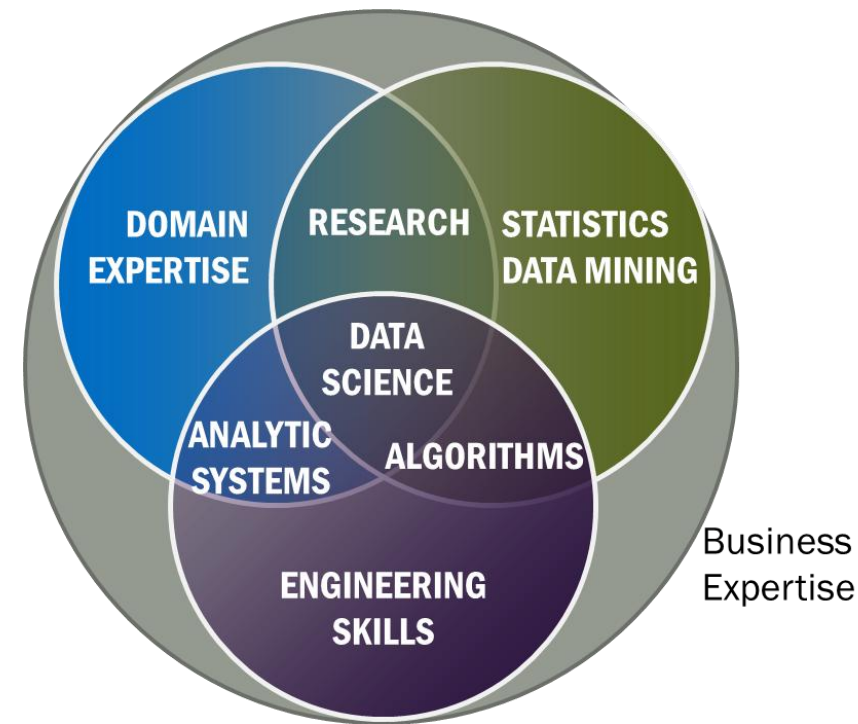


WPs interaction and main tasks



EDISON Approach (1): Data Science Definition

- Data Science and Data Scientist Professional (DSP) definition
 - Including organizational roles and general competence domains
 - General Data Science literacy – common knowledge to use Data Science tools and specify tasks for DSP/engineers

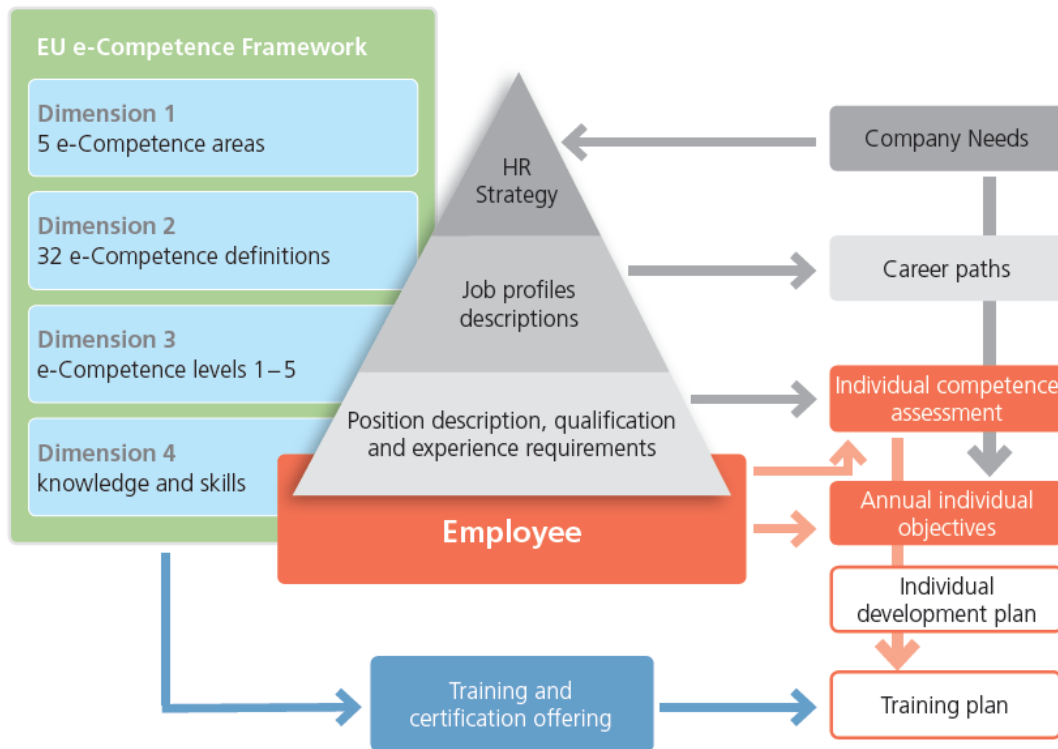


Definition by NIST Big Data WG (2014-2015)

*A **Data Scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of expertise in business needs, domain knowledge, analytical skills, and programming and systems engineering expertise to manage the end-to-end scientific method process through each stage in the **big data lifecycle**.*

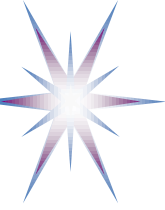
EDISON Approach (2): e-CFv3.0 and CF-DS

- Competence Framework for Data Science (CF-DS) definition will be built based on European e-Competence framework for IT (e-CFv3.0)
 - Linking scientific research lifecycle, organizational roles, competences, skills and knowledge
 - Defining Data Science Body of Knowledge (DS-BoK)
 - Mapping CF-DS and DS-BoK to academic disciplines



- Multiple use of e-CFv3.0 within ICT organisations
- Provides basis for individual career path, competence assessment, training and certification

- EDISON CF-DS will be used for defining DS-BoK and MC-DS, linking organizational functions and required knowledge
- Provide basis for individual (self) training and certification

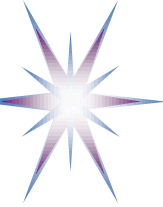


e-CFv3.0 Internal Structure: Refactoring for CF-DS

European e-Competence Framework 3.0 overview

Dimension 1 5 e-CF areas (A – E)	Dimension 2 40 e-Competences identified	Dimension 3 e-Competence proficiency levels e-1 to e-5, related to EQF levels 3–8				
		e-1	e-2	e-3	e-4	e-5
A. PLAN	A.1. IS and Business Strategy Alignment					
	A.2. Service Level Management					
	A.3. Business Plan Development					
	A.4. Product/Service Planning					
	A.5. Architecture Design					
	A.6. Application Design					
	A.7. Technology Trend Monitoring					
	A.8. Sustainable Development					
	A.9. Innovating					
B. BUILD	B.1. Application Development					
	B.2. Component Integration					
	B.3. Testing					
	B.4. Solution Deployment					
	B.5. Documentation Production					
	B.6. Systems Engineering					
C. RUN	C.1. User Support					
	C.2. Change Support					
	C.3. Service Delivery					
	C.4. Problem Management					
D. ENABLE	D.1. Information Security Strategy Development					
	D.2. ICT Quality Strategy Development					
	D.3. Education and Training Provision					
	D.4. Purchasing					
	D.5. Sales Proposal Development					
	D.6. Channel Management					
	D.7. Sales Management					
	D.8. Contract Management					
	D.9. Personnel Development					
	D.10. Information and Knowledge Management					
	D.11. Needs Identification					
	D.12. Digital Marketing					
E. MANAGE	E.1. Forecast Development					
	E.2. Project and Portfolio Management					

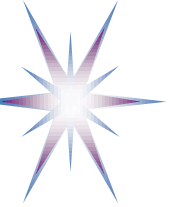
- 4 Dimensions
 - Competence Areas
 - Competences
 - Proficiency levels
 - Skills and Knowledge
- 5 Competence Area defined by ICT Business Process stages
 - Plan
 - Build
 - Deploy
 - Run
 - Manage
- > Refactor to Scientific Research (or Scientific Data) Lifecycle
 - See example of RI manager at IG-ETRD wiki and meeting
- Each competence has 5 proficiency level
 - Ranging from technical to engineering to management to strategist/expert level
- Knowledge and skills property are defined for/by each competence and proficiency level (not unique)



EDISON Difference – Long Term Perspective

- Main project results are targeted for “durable” use and future effect
 - Education and Data Science curriculum innovation to take effect in 3-5 years
 - Training and re-skilling to solve short-term needs
 - Sustainability model and stakeholder involvement
- Propose multiple paths for Data Science career
 - Historical career cycle was 10-15 years
 - Now 5-7 year in current conditions of accelerated technologies change
 - Need to understand trends
 - USA PCAST report (August 2015) on Networking and Information Technology R&D (NITRD) stresses the importance of education and governmental support [ref]
- Community involvement and contribution
 - EDISON Liaison Groups of experts
 - Champion universities and organisations

[ref] <https://www.whitehouse.gov/blog/2015/08/07/pcast-assesses-federal-information-technology-rd>



Topics for Discussion

e-Infrastructures for Open Scholarship: Interoperability and Ecosystem/Sustainability and Advancement

- How to deconstruct Open Scholarship into basic ideas/elements and e-infra-matching elements? The goal is to agree on terminology, goals and objectives, and perhaps more importantly priorities.
 - (1) Terminology is important, especially in new areas such as Data Science or Big Data
 - (2) Define goal and expected outcome – future vision and long term trends
 - (3) Innovation from Science to Industry – closer relations with industry, also as a possible career path for researchers (and Data Scientists in particular)
- How can e-Infrastructures facilitate Open Scholarship? The focus should be on services (including training and support), costs, incentives for use, barriers and how to go about solving them.
 - (1) Address human factor: Education and Training infrastructure as a part of and linked to e-Infrastructure
 - (2) Education and training linked to Scientific research lifecycle (also Research Data lifecycle) and organizational roles
- Open Scholarship requires transparency at all levels of the research life-cycle, which effectively leads to trust and uptake. What/who do we need to mobilize the right stakeholders to build this trust?
 - (1) Trustworthiness (trust by design) and data-centric security model
 - (2) Federated infrastructure – to increase trust by collective management of policies
 - (3) Privacy and Opacity (as opposite to transparency)
 - (4) Ethics and Responsible use: Data based economy == Resource based economy