6.20 Assistance to the development of the taxonomy strategy for the relevant data catalogues

Document Control Information

Settings	Value
Document Title:	D6.20 Assistance to the development of the taxonomy strategy for the relevant data catalogues
Project Title:	ExPaNDS
Document Author:	Isabelle Boscaro-Clarke, Lead WP6 (Diamond)
Project Coordinator:	Patrick Fuhrmann (DESY)
Doc. Version:	1
Dissemination level:	Public
Date:	31/08/2021

Table of Contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

1. Abstract

National facilities have tended to develop their own subject area taxonomies, which are usually applied to peer-reviewed publications. These subject area taxonomies have been developed independently from those used to catalogue data. An earlier deliverable[1] identified that the Subject Resource Application Ontology (SRAO) should be used and extended as the common ExPaNDS Taxonomy. Some facilities had existing baseline taxonomies[2] for both subject areas and techniques, which we shared in our previous deliverable "D6.19 Assistance to the development of taxonomy strategy for the relevant data catalogues". A benchmarking exercise against a range of vocabularies provided an opportunity to highlight areas for improvement and raise a general awareness of how well these vocabularies compare to each other.

2. Methodology

A qualified librarian carried out a benchmarking of all terminologies against the following existing catalogues:

- Subject Resource Application Ontology (SRAO)[3]
- European Science Vocabulary (EuroSciVoc) taxonomy/thesaurus[4]
- The Organisation for Economic Co-operation and Development (OECD) subject vocabulary[5]
- PaN Subject and Techniques published[2] (deliverable D6.19)

3. Outcomes

Subject Resource Application Ontology (SRAO)

When looking at the SRAO taxonomy and comparing it with our previous ExPaNDS deliverable (D6.19), we found that 78 terms out of 426 have an equivalent term or keyword.

When comparing both sets of terms we also found that 45 terms out of the 125 from the D6.19 deliverable taxonomy have no equivalent in the SRAO taxonomy and would ideally need to be added to be of use for the ExPaNDS communities. Some key areas are missing like energy and their subcategories, quantum materials and their subcategories and catalysis.

It should be noted that the fields of studies of Photon and Neutron (PaN) facilities are narrower than the fields present in the complete SRAO taxonomy, and as such the large number of missing terms from the taxonomy is expected as only relevant subjects are included on a facility's taxonomy.

[1] https://zenodo.org/record/4806026#.YSNus4hKiUk
[2] https://zenodo.org/record/4715000#.YSOBt4hKhPY
[3] https://fairsharing.org/bsg-s001177/
[4] https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/euroscivoc
[5] https://www.oecd.org/science/inno/38235147.pdf

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

European Science Vocabulary (EuroSciVoc)

The EuroSciVoc taxonomy consists of a very detailed list of terms (991 terms), it has 125 terms in common with the taxonomy (45 keywords), but also contains a number of very specific terms which are too specific to describe a science subject and are more akin to keywords.

The Organisation for Economic Co-operation and Development (OECD) subject vocabulary

The OECD subject vocabulary is more general than the EuroSciVoc taxonomy and is more suited to the context of ExPaNDS as it describes fields of science and technologies. It has 60 terms in common with the taxonomy (2 Keywords).

4. Conclusions

While both the OECD and SRAO taxonomies could be used as a strong foundation to the ExPaNDS taxonomy, it was concluded that the SRAO subject list was more complete and more suitable for the PaN community.

To be used as the common ExPaNDS Taxonomy, however the SRAO will need to be enhanced with the missing subjects. Our recommendation would be that the ExPaNDS Work Package 3 evaluates the list of subjects that have been identified and submits their list of potential additions to the SRAO for review and consideration. A previously submitted deliverable (D3.2[6]) suggests extending the SRAO taxonomy by proposing new terms or entire branches via the SRAO GitHub issues mechanism which is maintained and supported by the FAIRsharing[7] team.

The list of missing subjects that we have identified as part of the benchmarking exercise and the comparisons can be found in the numbered appendices of this document below – 5.1 to 5.4.

In addition, we further recommend that the chosen taxonomy is managed carefully and should remain stable before it is used across the ExPaNDS PaN facilities and respective data catalogues. Additions and changes to the taxonomy will have to be applied retrospectively to the data, in order to benefit from the structured taxonomy. This would be very time consuming and convoluted to do across all data catalogues.

5. Appendices

5.1 SRAO comparison:

Please see below the comparison report between the SRAO subject list and the taxonomy previously submitted as part of D6.19. Please note that only equivalent terms are listed below, and exact vocabulary matches are identified with a green background; the other lines may be synonyms or slight variations.

[6] https://zenodo.org/record/4806026
[7] https://fairsharing.org/bsg-s001177/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

SRAO Term	Taxonomy submitted as part of deliverable D6.19
Aerospace Engineering	Aerospace
Agriculture	Agriculture & Fisheries
Ancient Cultures	Cultural Heritage (?)
Animal Nutrition	
Anthropology	
Artificial Intelligence	Artificial Intelligence
Astrophysics and Astronomy	Astronomy/Astrophysics/Astroparticles
Atmospheric Science	Atmospheric Processes
Atomic, Molecular, Optical and Plasma Physics	Matter under extreme conditions, warm dense matter, plasmas (?)
Biochemistry	Biochemistry
Biodiversity	Ecosystems & Biodiversity
Biology	Life Sciences & Biotech
Biomaterials	Biomaterials
Biophysics	Biophysics
Biotechnology	Biotechnology
Botany	Plant science
Chemical Engineering	Chemical Engineering
Chemistry	Chemistry
Composite Materials	Composite Materials
Computer Science	Information & Communication Technologies
Condensed Matter Physics	Hard condensed matter - electronic properties
Condensed Matter Physics	Hard condensed matter - structures
Dentistry	Dentistry
Drug Discovery	Drug Discovery
Earth Science	Earth Sciences & Environment
Ecology	Ecosystems & Biodiversity
Ecosystem Science	Ecosystems & Biodiversity
Engineering Science	Engineering & Technology
Environmental Science	Earth Sciences & Environment
Epigenetics	
Evolutionary Biology	Evolutionary science (Also Evolution)
Fine Arts	
Fisheries Science	Agriculture & Fisheries
Food Chemistry	Food Science (?)
Food Process Engineering	Food Science (?)
Genetics	Genetics
Geochemistry	Geochemistry
Geology	Geology
Geophysics	Geophysics

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Health Science	Health & Wellbeing
Human Genetics	Genetics
Humanities	Humanities
Hydrogeology	Water sciences/Hydrology
Hydrology	Water sciences/Hydrology
Industrial Engineering	Industrial Engineering
Infectious Disease Medicine	Infectious Diseases
Life Science	Life Sciences & Biotech
Marine Biology	Marine science/Oceanography
Materials Engineering	Material Sciences
Materials Engineering	Materials Engineering & Processes
Mathematics	Mathematics
Metallurgy	Metallurgy
Mineralogy	Mineralogy
Molecular Chemistry	Molecular Complexes (?)
Molecular Physical Chemistry	Molecular Physics
Nanotechnology	Nanoscience/Nanotechnology
Neurology	Neurology
Neuroscience	Neurology (?)
Oceanography	Marine science/Oceanography
Oncology	Cancer
Ophthalmology	Ophthalmology
Organic Chemistry	Organic Chemistry
Paleontology	Paleontology
Parasitology	Parasitology
Physical Chemistry	Physical Chemistry
Physics	Physics
Plant Cultivation	Plant science (?)
Polymer Chemistry	Polymer Science
Polymer Physics	Polymer Science
Polymer Research	Polymer Science
Soft Matter Physics	Soft condensed matter physics
Software Engineering	Computing & software technologies
Structural Biology	Structural Biology
Surface Science	Surfaces
Theoretical Chemistry	Theoretical Chemistry
Veterinary Medicine	Veterinary Medicine
Virology	
Water Research	Water sciences/Hydrology

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

5.2 SRAO missing terms:

Antibiotic Resistance
Autoimmune Diseases
Automotive
Carcinogens
Catalysis
Ceramics
Climate Change
Communication & Networks
Components & Micro-systems
Corrosion
Data management / presentation
Data processing
Desertification & Pollution
Detectors
Disease in the Developing World
Drug Delivery
Electronics
Energy
Energy Materials
Energy Storage
Ferroelectrics
High energy & particle physics
Magnetism
Metal-Organic Frameworks
Metrology
Multiferroics
Natural disaster
Neurodegenerative Diseases
Non-Communicable Diseases
Nuclear Waste
Optics
Organometallic Chemistry
Osteoporosis
Pathogens
Perovskites
Planetary Geology
Quantum Materials
Radioactive Materials
Superconductors
Sustainable Energy Systems

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Technique Development
Theoretical Physics
Thermoelectrics
Vaccines
Zeolites

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

5.3 EuroSciVoc comparison:

Please see below the comparison report between the EuroSciVoc subject list and the taxonomy previously submitted as part of D6.19. Please note that only equivalent terms are listed below, and exact vocabulary matches are identified with a green background; the other lines may be synonyms or slight variations.

EuroSciVoc Taxonomy Term	Taxonomy submitted as part of deliverable D6.19	Keyword
additive manufacturing		Additive Manufacturing
agricultural sciences	Agriculture & Fisheries (?)	
agriculture	Agriculture & Fisheries	
agriculture, forestry, and fisheries	Agriculture & Fisheries	
alzheimer		Alzheimer's Disease
amyotrophic lateral sclerosis		Motor Neurone Disease
Animal Feed		Animal Feed
antibiotic resistance	Antibiotic Resistance	
art history	Cultural Heritage (?)	
artificial intelligence	Artificial Intelligence	
asteroids		Asteroids
astrophysics	Astronomy/Astrophysics/Astroparticles	
atmospheric sciences	Atmospheric Processes	
autoimmune diseases	Autoimmune Diseases	
automotive engineering		Automotive
bacteriology		Bacteria
biochemistry	Biochemistry	
biofuels		Biofuel
bioleaching		Bioleaching
biological sciences	Life Sciences & Biotech (?)	
biomaterials	Biomaterials	
biophysics	Biophysics	
bioremediation		Bioremediation

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

botany (photosynthesis)		Photosynthesis
botany (plant sciences)	Plant science	
breast cancer		Breast Cancer
cancer	Cancer	
carbon capture engineering		Carbon Capture and Storage (CCS)
cardiovascular diseases		Cardiovascular Disease
ceramics	Ceramics	
chemical engineering	Chemical Engineering	
chemical sciences	Chemistry (?)	
climatic changes	Climate Change	
coating and films	Thin films	
colorectal cancer		Colorectal Cancer
computer and information sciences	Information & Communication Technologies	
condensed matter physics	Hard condensed matter - electronic properties	
condensed matter physics	Hard condensed matter - structures	
condensed matter physics	Soft condensed matter physics	
coronavirus		COVID-19
dairy		Dairy Products
data processing	Data Processing	
deep learning	Artificial Intelligence	
diabetes		Diabetes
ecosystems	Ecosystems & Biodiversity	
electric batteries		Batteries
electric vehicles		Electric Vehicles
electrolysis (corrosion)	Corrosion	
energy and fuels	Energy	
environmental sciences	Earth Sciences & Environment (?)	
enzymes		Enzymes
epigenetics		Epigenetics
epilepsy		Epilepsy

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

evolutionary biology	Evolutionary science	
fertility		Infertility (probably should've been Fertilty in reflection)
fisheries	Agriculture & Fisheries	
food and beverages	Food Science	
fuel cell		Fuel Cells
genetics and heredity	Genetics	
geochemistry	Geochemistry	
geology	Geology	
geophysics	Geophysics	
health sciences	Health & Wellbeing (?)	
hepatitis B		Hepatitis B Virus (HBV)
hepatitis C		Hepatitis C Virus (HCV)
history and archaeology	Cultural Heritage	
hiv		Human Immunodeficiency Virus (HIV)
hydrogeology	Water sciences/Hydrology	
immunotherapy		Immunotherapy
infectious disease	Infectious Diseases	
influenza		Influenza
inorganic chemistry	Inorganic Chemistry	
leukemia		Leukaemia
malaria		Malaria
manufacturing engineering	Industrial Engineering	
marine biology	Marine science/Oceanography	
materials engineering	Materials Engineering & Processes	
mechanical engineering	Industrial Engineering (?)	
medical biotechnology	Biotechnology	
medical engineering	Biotechnology (?)	
medical genetics	Genetics	
metallurgy	Metallurgy	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Meteorites		Meteorites
mineralogy	Mineralogy	
molecular and chemical physics	Molecular Physics	
nanotechnology	Nanoscience/Nanotechnology	
natural disaster	Natural disaster, Desertification & Pollution	
neurology	Neurology	
nuclear waste management	Nuclear Waste	
oceanography	Marine science/Oceanography	
odontology	Dentistry	
optics	Optics	
organic chemistry	Organic Chemistry	
orthodontics	Dentistry	
paediatric cardiology (congenital heart disease)		Congenital Heart Defects (CHD)
palaeontology	Palaeontology	
paleoanthropology		Paleoanthropology
parasitology	Parasitology	
parkinson		Parkinson's Disease
particle physics	High energy & particle physics	
periodontics	Dentistry	
pharmaceutical drug	Drug Discovery & Drug Delivery (?)	
pharmaceutical drug	Drug Discovery & Drug Delivery (?)	
pharmacology and pharmacy	Drug Discovery & Drug Delivery (?)	
photocatalysis		Photocatalysis
physical chemistry	Physical Chemistry	
planetary geology	Planetary Geology	
polymer science	Polymer Science	
prostate cancer		Prostate Cancer
radiation chemistry	Radioactive Materials (?)	
renewable energy	Sustainable Energy Systems	
semiconductor		Semiconductors

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

skin cancer		Skin Cancer
soft matter physics	Soft condensed matter physics	
software	Computing & software technologies	
solar energy (photovoltaic)		Photovoltaics
spintronics		Spintronics
structural biology	Structural Biology	
superconductor		Superconductors
theoretical physics	Theoretical Physics	
tropical medicine	Disease in the Developing World	
tuberculosis		Tuberculosis (TB)
vaccines	Vaccines	
volcanology		Volcanology

5.4 OECD comparison:

Please see below the comparison report between the OECD subject list and the taxonomy previously submitted as part of D6.19. Please note that only equivalent terms are listed below, and exact vocabulary matches are identified with a green background; the other lines may be synonyms or slight variations.

OECD Term	**Taxonomy submitted as part of deliverable D6.19**	**Keyword**
Aerospace engineering	Aerospace	
Agriculture, Forestry, and Fisheries	Agriculture & Fisheries	
Allergy	Allergic Diseases	
Astronomy	Astronomy/Astrophysics/Astroparticles	
Atomic, molecular and chemical physics	Molecular Physics	
Biochemistry and molecular biology	Biochemistry	
biofuels		Biofuel
Biomaterials (as related to medical implants, devices, sensors)	Biomaterials	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Biophysics	Biophysics	
Ceramics	Ceramics	
Chemical engineering	Chemical Engineering	
climatic research	Climate Change	
Clinical neurology	Neurology	
Coating and films	Surfaces, interfaces and thin films (?)	
Composites (including laminates, reinforced plastics, cermets, combined natural and synthetic fibre fabrics; filled composites)	Composites	
Computer and information sciences	Information & Communication Technologies	
Condensed matter physics	Hard condensed matter - electronic properties	
Condensed matter physics	Hard condensed matter - structures	
Condensed matter physics	Soft condensed matter physics	
Dentistry, oral surgery and medicine	Dentistry	
Earth and related Environmental sciences	Earth Sciences & Environment	
Energy and fuels	Energy	
Evolutionary biology	Evolutionary science	
Fluids and plasma physics (including surface physics)	Matter under extreme conditions, warm dense matter, plasmas (?)	
Fluids and plasma physics (including surface physics)	Surfaces, interfaces and thin films (?)	
Food and beverages	Food Science	
Genetics and heredity	Genetics	
Geochemistry and geophysics	Geochemistry	
Geochemistry and geophysics	Geophysics	
Geology	Geology	
Health-related biotechnology	Biotech & Biological Systems or Biotechnology	
History and Archaeology	Cultural Heritage	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Infectious diseases	Infectious Diseases	
Inorganic and nuclear chemistry	Inorganic Chemistry	
Marine biology, freshwater biology, limnology; Ecology; Biodiversity conservation	Marine science/Oceanography	
Materials engineering	Materials Engineering & Processes	
Mathematics	Mathematics	
Mechanical engineering	Industrial Engineering (?)	
Meteorology and atmospheric sciences	Atmospheric Processes	
Mineralogy	Mineralogy	
Nano-technology	Nanoscience/Nanotechnology	
Neurosciences (including psychophysiology)	Neurology	
Oceanography, Hydrology, Water resources	Marine science/Oceanography	
Oceanography, Hydrology, Water resources	Water sciences/Hydrology	
Oncology	Cancer	
Ophthalmology	Ophthalmology	
Optics	Optics	
Organic chemistry	Organic Chemistry	
Palaeontology	Palaeontology	
Parasitology	Parasitology	
Particles and fields physics	High energy & particle physics	
Pharmacology and pharmacy	Drug Discovery & Drug Delivery (?)	
Physical chemistry	Physical Chemistry	
Plant sciences, botany	Plant sciences	
Polymer science	Polymer Science	
Transport engineering	Transport	
Tropical medicine	Disease in the Developing World (?)	
Veterinary science	Veterinary Medicine	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857641.

Volcanology		Volcanology