WORKING GROUPS REPORT

Sofia, BU, 12-13 March 2019

WG4 WORKSHOP: Data storage and archiving strategies

1. Attendance

a. WG4 Leader and Co-Leader: Dagmar Triebel and Gila Kahila Bar-Gal
b. Convener/Rapporteur: Hilary Goodson
c. Number of Participants: 14
d. Inclusiveness rate of the WG participation:
   - Gender balance M/F: 8/6
   - Geographical distribution: Belgium, Bulgaria, France, Germany, Israel, Luxembourg, Netherlands, Norway, Spain.
   - Seniority: 5 early-career investigators

2. Objectives of the WG workshop

The objectives of the workshop were to (1) Identifying general strategies in the long-term storage and archiving of object data and assigned multimedia files. This include the discussion on technical and content standards as well as specific conditions which might arise by legal and administrative requirements. (2) Identifying the best suitable approach for storage and archiving solutions for biological and earth science collections (large and small sized collections). The discussion was based on detailed description of various archiving systems used by hosting agencies and research data infrastructure organisations. (3) The relations among major European open science and research data infrastructures – benefits.

3. Topics discussed

Definition of “data archiving” and related terms and the difference from data backup. What needs to be archived and for how long? The emphasis on biological scientific collection versus “natural history” collections including earth science objects.

Methods that stage the primary data itself to a cheaper storage media with quick data retrieval. What are the specific needs in archiving data assigned to physical (museum) objects? What about digital objects (data) where the physical objects are gone (e.g., certain sequence data) or never existed as museum object (observation data).

Storage concepts and models used: Presentations of different models and solutions pros and cons for each one. The Open Archival Information System reference model was raised several times. The group collected and listed useful links and material for major concepts and (community) standards on data storage and technical solutions on (FAIR) data archiving and long-term preservation (see under https://costmobilise.biowikifarm.net/wiki/Useful_links_and_materials).

The general discussions proceed on the history of data archiving standards and standardisation, principles of archival of digital assets, collection-relevant archive formats, collection-related metadata standards and norms and gaps in standards relevant for collection data archiving.

Practice in archiving large (multimedia) files and other “big amount of data”, e.g. on specific concepts to store and archive: (a) OTU (taxon)-defining voucher-less molecular datasets (from molecular phylogenetic studies) and (b) meta-omics datasets from environmental samples (from community barcoding approaches).

Case studies from two large natural history museums and from sequencing platforms were presented. A perspective view on the application of ontologies for facilitation of reasonable access to archived biological data was presented.

Summary: During WG4 workshop 10 presentations were given and can be downloaded from the WG4 Sofia Workshop website.

In reference to completion of Y1 objectives we made a landscape analysis and identified existing communities and expertise. We have now a large group of more than 25 persons interested in WG4 listed as “contributors” in the OSF platform (see under https://osf.io/wq7ej/). They are mainly members of related work packages of large collection-related EU projects and ready to give advices. A smaller group of 14-16 experts from the stakeholder groups, service providers and target group are willing to cooperate and work together as core group.

For Y2 we plan a number of telco activities and actions as well as one face-to-face workshop which we will document.

4. Major outcomes

Several important statements:

- We have to define timeframes of archiving: 10 years to infinity.
- The parameters scalability, reliability and reachability define our scope.
- We preserve and archive to get access (Reconfirmation: access is confirmation of preservation worthiness, verification).
- Archiving practices should result in FAIR data.
- Archiving of data is migration of data.

5. Challenges faced for Y2
   - Complexity of archiving processes and factors in general

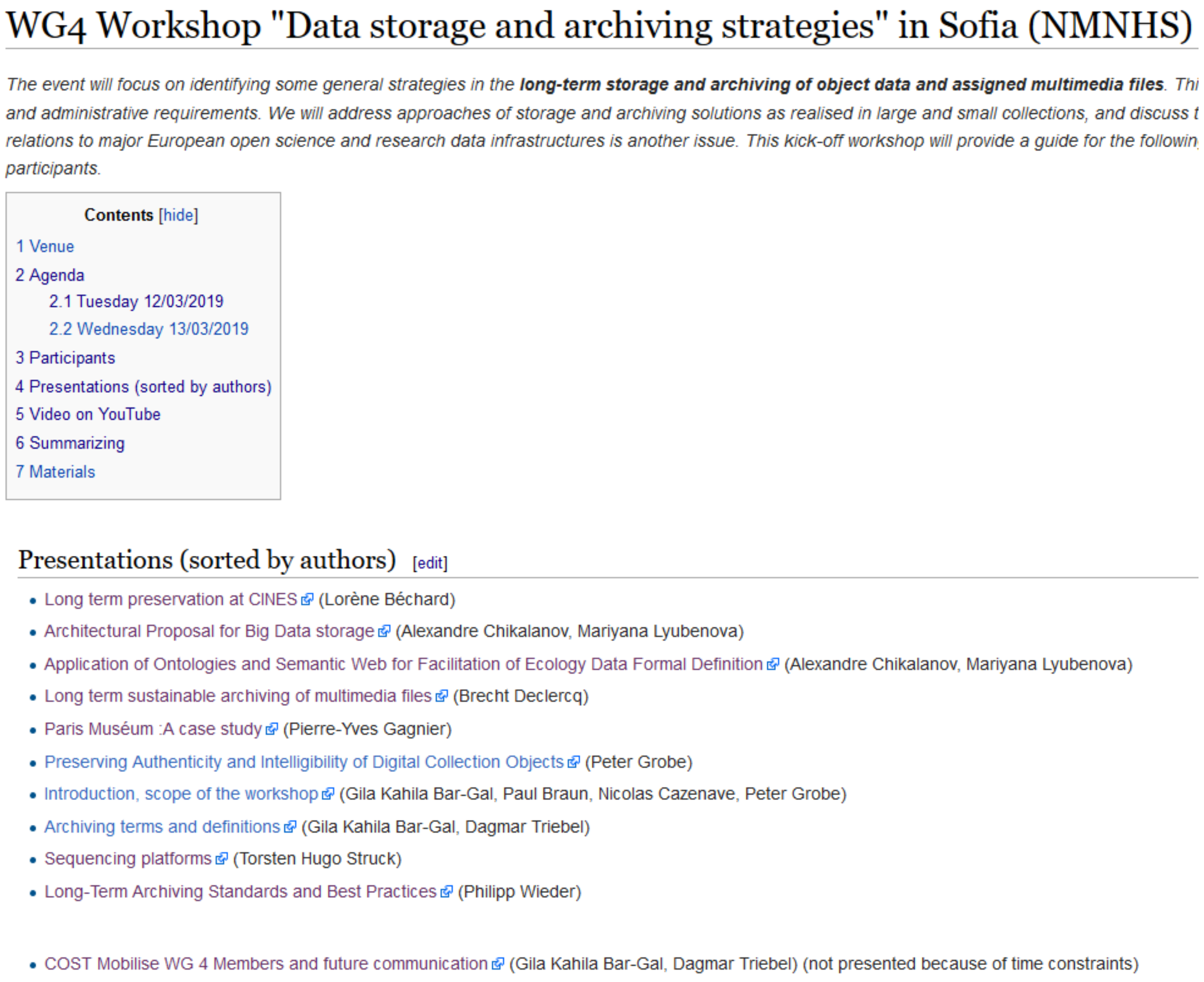
- Focussing on requests of target group “scientific collections”
- Within the framework of OAIS standards, our focus is on AIP, always with thought to alignment with SIP and DIP as defined within the community areas/ working groups
- Agreement on sets of best practices to archive data and enable FAIR usage

6. Next steps

Starting two online documentations:

- a wiki page for definitions of core terms
- a GoogleDoc to write a leaflet with first (core) recommendations on data archiving in natural history collections

ANNEXES: Screenshots from online information on the WG4 Workshop



WG4 Workshop "Data storage and archiving strategies" in Sofia (NMNHS)

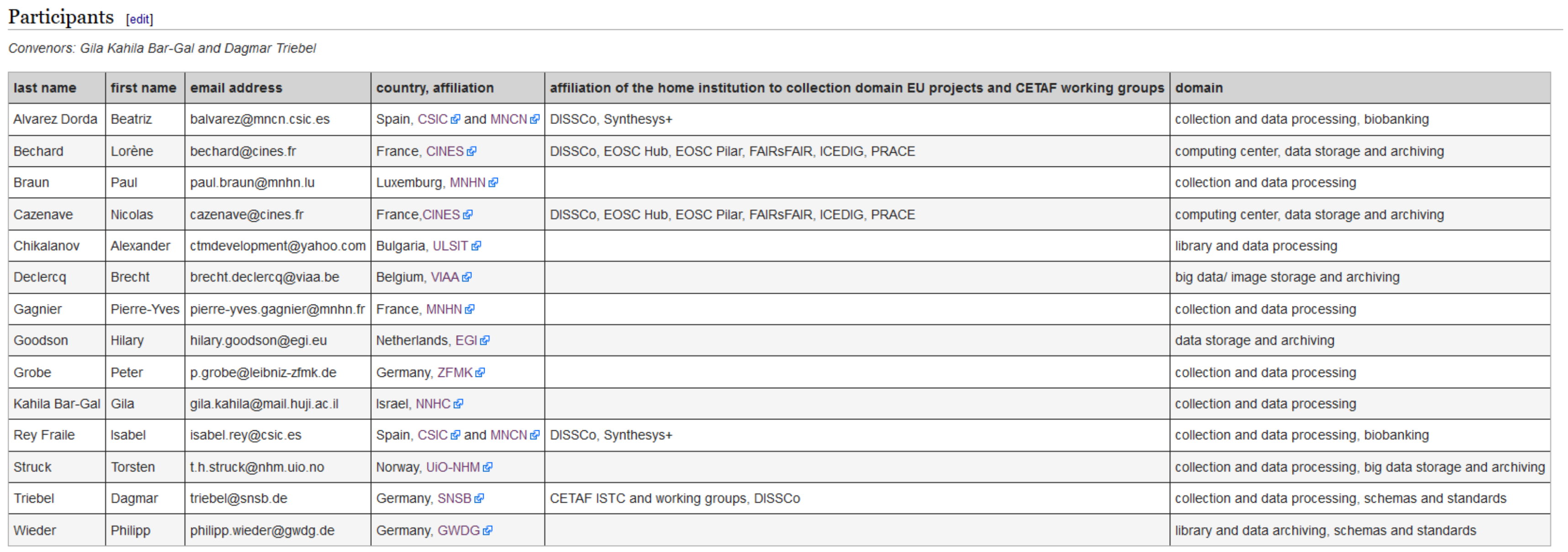
*The event will focus on identifying some general strategies in the **long-term storage and archiving of object data and assigned multimedia files**. Thi and administrative requirements. We will address approaches of storage and archiving solutions as realised in large and small collections, and discuss t relations to major European open science and research data infrastructures is another issue. This kick-off workshop will provide a guide for the followin, participants.*

Contents [hide]

Presentations (sorted by authors) [edit]

- Long term preservation at CINES (Lorène Béchard)
- Architectural Proposal for Big Data storage (Alexandre Chikalanov, Mariyana Lyubenova)
- Application of Ontologies and Semantic Web for Facilitation of Ecology Data Formal Definition (Alexandre Chikalanov, Mariyana Lyubenova)
- Long term sustainable archiving of multimedia files (Brecht Declercq)
- Paris Muséum :A case study (Pierre-Yves Gagnier)
- Preserving Authenticity and Intelligibility of Digital Collection Objects (Peter Grobe)
- Introduction, scope of the workshop (Gila Kahila Bar-Gal, Paul Braun, Nicolas Cazenave, Peter Grobe)
- Archiving terms and definitions (Gila Kahila Bar-Gal, Dagmar Triebel)
- Sequencing platforms (Torsten Hugo Struck)
- Long-Term Archiving Standards and Best Practices (Philipp Wieder)

- COST Mobilise WG 4 Members and future communication (Gila Kahila Bar-Gal, Dagmar Triebel) (not presented because of time constraints)



Participants [edit]

Convenors: Gila Kahila Bar-Gal and Dagmar Triebel

last name	first name	email address	country, affiliation	affiliation of the home institution to collection domain EU projects and CETAF working groups	domain
Alvarez Dorda	Beatriz	balvarez@mncn.csic.es	Spain, CSIC and MNCN	DISSCo, Synthesys+	collection and data processing, biobanking
Bechard	Lorène	bechard@cines.fr	France, CINES	DISSCo, EOSC Hub, EOSC Pilar, FAIRsFAIR, ICEDIG, PRACE	computing center, data storage and archiving
Braun	Paul	paul.braun@mnhn.lu	Luxemburg, MNHN		collection and data processing
Cazenave	Nicolas	cazenave@cines.fr	France,CINES	DISSCo, EOSC Hub, EOSC Pilar, FAIRsFAIR, ICEDIG, PRACE	computing center, data storage and archiving
Chikalanov	Alexander	ctmdevelopment@yahoo.com	Bulgaria, ULSIT		library and data processing
Declercq	Brecht	brecht.declercq@viaa.be	Belgium, VIAA		big data/ image storage and archiving
Gagnier	Pierre-Yves	pierre-yves.gagnier@mnhn.fr	France, MNHN		collection and data processing
Goodson	Hilary	hilary.goodson@egi.eu	Netherlands, EGI		data storage and archiving
Grobe	Peter	p.grobe@leibniz-zfmk.de	Germany, ZFMK		collection and data processing
Kahila Bar-Gal	Gila	gila.kahila@mail.huji.ac.il	Israel, NNHC		collection and data processing
Rey Fraile	Isabel	isabel.rey@csic.es	Spain, CSIC and MNCN	DISSCo, Synthesys+	collection and data processing, biobanking
Struck	Torsten	t.h.struck@nhm.uio.no	Norway, UiO-NHM		collection and data processing, big data storage and archiving
Triebel	Dagmar	triebel@snsb.de	Germany, SNSB	CETAF ISTC and working groups, DISSCo	collection and data processing, schemas and standards
Wieder	Philipp	philipp.wieder@gwdg.de	Germany, GWDG		library and data archiving, schemas and standards