TEI guidelines: born to be open

Laurent Romary

The TEI guidelines: born to be open

Laurent Romary, Inria team ALMAnaCH

Why this? Why now?

- Very strong institutional demand with regards to open science
  - With hype keywords replacing sometimes scholarly considerations
- Identifying where we are as a community (“digital scholarship”)
  - Where is this feeling of being kind-of forerunners coming from?
- Setting up a coherent picture
  - Publications – data (“sources”)
- Three personal sources of inspiration
  - 15 years of open access activism
    - Examples from the French setting
  - 15 years in setting up DARIAH…
    - Open science success stories (I am not even mentioning all of them – e.g. *Open Methods*)
  - 27 years in the TEI business
    - Demonstrating that the TEI is the perfect place to be open

Ouvrir la science…

Why should science be (more) open?

- Fluidifying the circulation of scientific results
  - Confronting assumptions, validating them, reproducing them
  - Contributing to informed debates in our society
- Towards a systemic/systematic vision of open science
  - Opening everything, as much as possible, and at each and every stage of the research process
    - Project proposals
    - Rough and intermediary data
    - Research processes, models, formats (meta-science)
    - Results, analyses… and yes, publications
- Still, there is not one such single thing as “open science”
  - E.g. disciplines: technical, institutional or legal differences

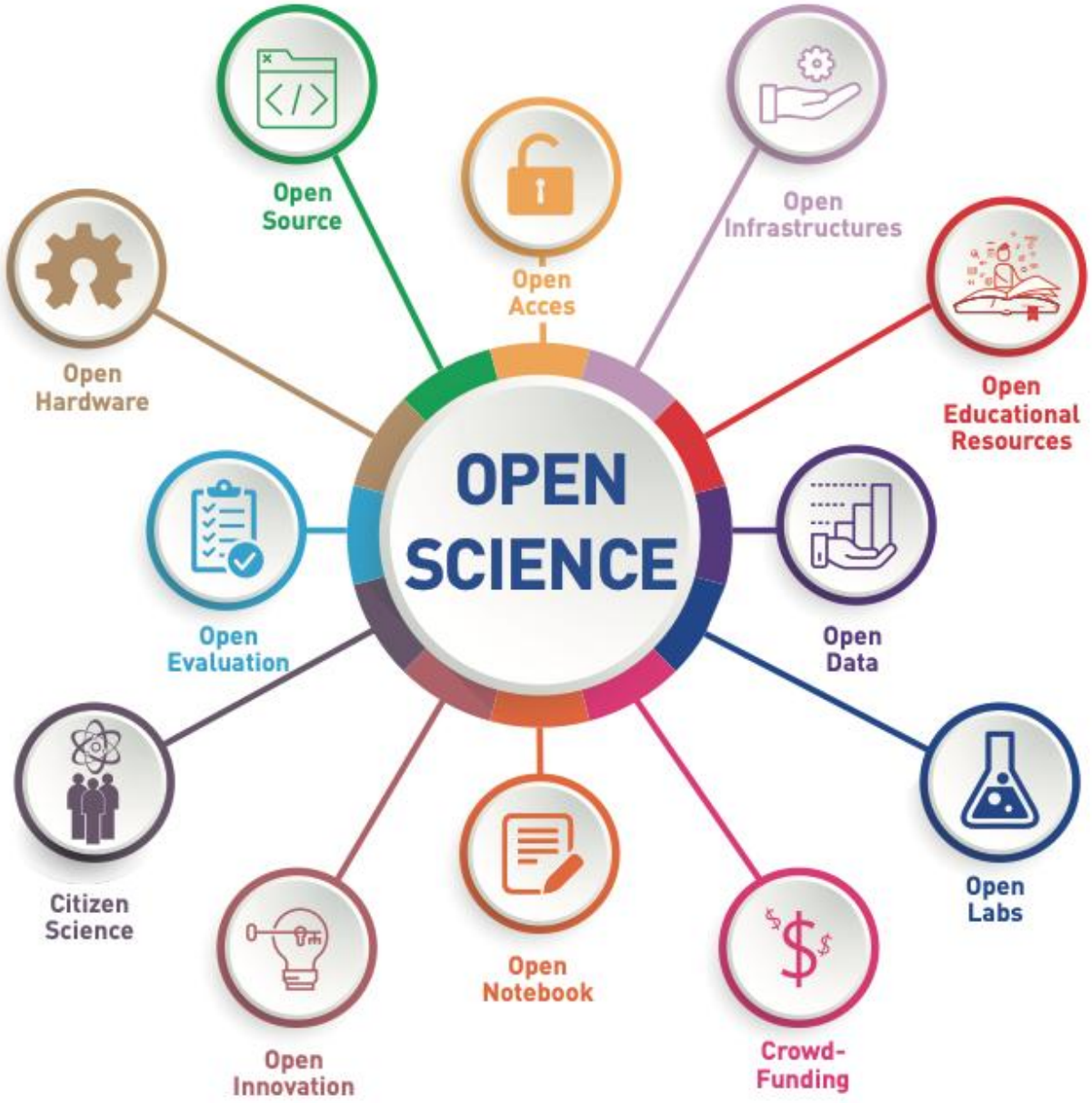
In memoriam: John Tennant

Is science not open yet?!?

- Paywalls, business models
  - Concentration of the publishing market

  …with the danger that the same happens to data in the future
- Reluctance to share, fear of plagiarism
  - “I’ll open later…”
- Lack of proper infrastructure
  - Standards, repositories, tools
- Lack of proper open science culture
  - Documenting, licencing, citing
- And not everything can be open
  - Copyright, personal information, sensitive data

The complexity of the open science landscape



Source: Towards a UNESCO Recommendation on Open Science — Building a Global Consensus on Open Science

How can science be more open ?

- Combining individual and collective initiatives
  - Adopting an early dissemination strategy for all the components of a research project
    - Open science should be part of the research methods themselves
  - Open science must be a component of the strategy of all HER institutions
    - Relation to reporting, assessment, communication
- Opening the debate on digital sovereignty
  - Who owns the content? Who is responsible for its curation? Who hosts the content?
  - Should not open science be based upon reliable and sustainable public infrastructures?
- Bringing open science on the political agenda
  - Having it part of governmental and funding agencies strategies
    - Funding requirements: cf. H2020 policies (open access, data management plans), and now ANR (Agence Nationale de la Recherche - French funding agency)
    - *Showing by doing*: ex. HCERES (French assessment agency) => https://hal-hceres.archives-ouvertes.fr
- Taking into account the legal context…

What does the law say? General principles

- Note: more precise information available from the DARIAH working group *Ethics and Legality in the Digital Arts and Humanities* (ELDAH)
- Research data must be open by default
  - Research data seen as public sector information
  - Responsibility: institution; operator: researcher
  - Specific (mandatory) cases: geographic or environmental data
- Possible exceptions
  - Intellectual property, personal data, strategic or transfer related information
  - Specific cases of cultural organisation: libraries, museums, archives
- European and French (sorry…) legal background
  - Loi n° 78-753 du 17 juillet 1978, CADA (Commission d'accès aux documents administratifs)
  - Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information (PSI)
    - Revised in 2013 (DIRECTIVE 2013/37/UE)
  - Loi dite Valter du 28 décembre 2015 relative à la gratuité et aux modalités de la réutilisation des informations du secteur public
  - Loi pour une République numérique - 7 octobre 2016
    - Integrated into the Code des relations entre le public et l'administration (livre III, titre II)

Open science in practice

Open science and open access

- *Old school open science*: open access to scholarly publications
  - An essential step before going deep sea
    - Simpler from an editorial and technical point of view
    - The baseline of scholarly communication
- Difficulties:
  - Access (paywalls)
  - Costs (subscriptions or Article processing charge - APCs – combined in hybrid journals)
  - Reuse – e.g. restrictions to text and data mining
  - Sovereignty: fragmented corpus of under-accessible, heterogeneous scholarly productions
    - Unavailable to researchers, communities, institutions, funders
    - Lack of sustainability
- Food for thought: example of the institutional policy at Inria
  - Deposit mandate in HAL (linked to annual reporting), including articles published in natively open journals (with APC)
  - Centralised budget for APCs – general ban of the hybrid model
    - OpenAPC
  - Un-subscribing from non-essential portfolios (Coquetry: Nature, Junk: Springer LNCS) and reinvestment in open platforms (e.g. Episciences)

But, but, but … why don’t you speak of the real open acces plaforms, Academia, ReasearchGate

- “One more word about ResearchGate/Academia.edu and why using these platforms will never be equal to proper self-archiving”
  - https://dariahopen.hypotheses.org/878
- “The Day I Removed my Publications from Academia & Research Gate”
  - https://icietla.hypotheses.org/114

- The central question: can you rely on non public services to deploy you open science policy?

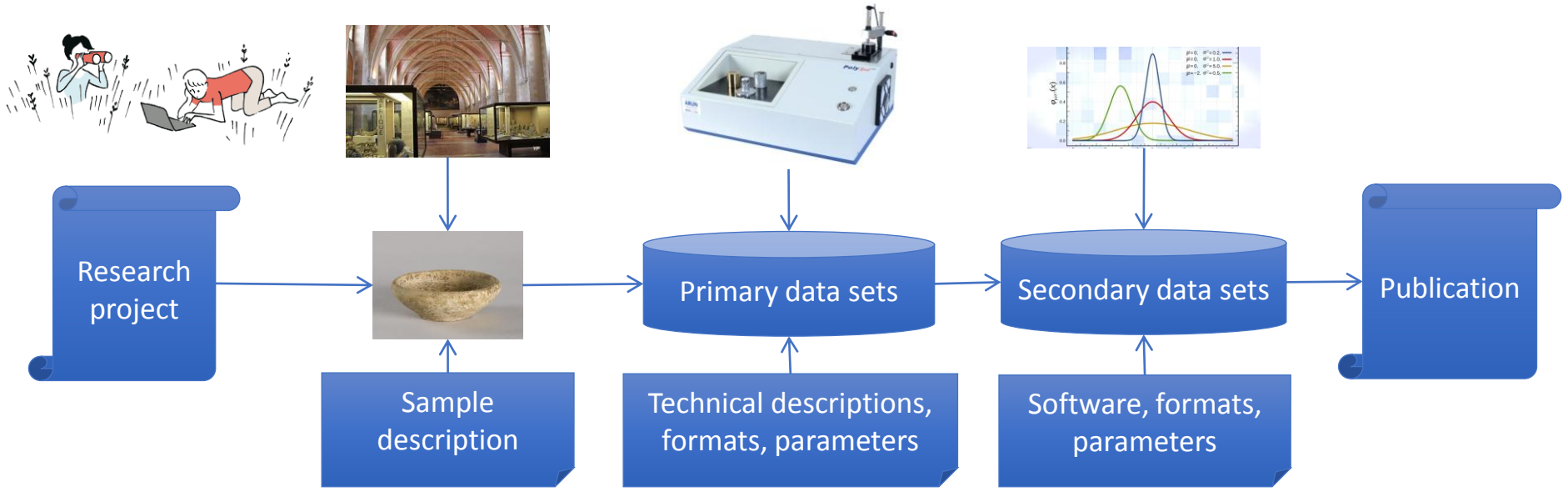
Open science and data management

- Opening/sharing research data is impossible without prior data management
  - Central questions related to data format, hosting, reuse conditions
- Mandatory deliverable by many funding agencies (e.g. ERC)

Dataset description, origin, size, datatypes and formats	Standardised formats and reference vocabulary
Associated metadata, identifiers (!)	Embargo, licence, data quality assurance
Identifying open and closed data sets, with explanations	Sustainability, procedures and resources for long-term availability (and/or protection)
Data repository	

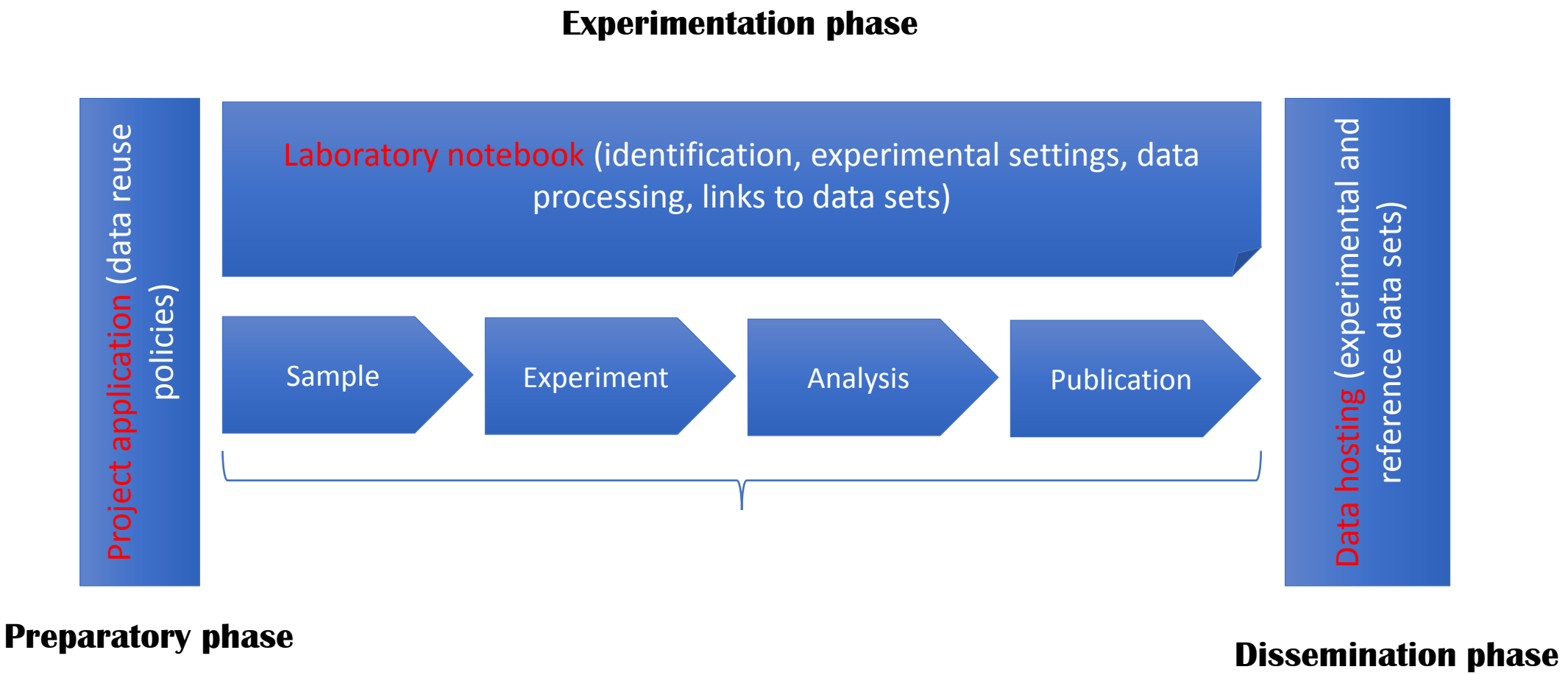
- Beyond DMPs – data management practices associated to research processes
  - Risk of seeing DMPs as administrative documents
  - Taking data management practices seriously, at our own service
- Real life example(s): dataset life cycle in heritage science

Beyond DMPs: tracing data sets in heritage science



Work carried out in the context of the Dopamine project, funded by the DIM MAP network (région Ile de France)

Towards a virtuous environment for experimental data management (and more)



Action points at each stage of the research process

- Preparatory phase
  - Project proposal (funding proposal, access request to the facility)
  - Data management plan
  - Anticipating on the reuse conditions: hand-shake between the different stakeholders
    - The data reuse charter as a central instrument
- Experimentation phase
  - Digital laboratory notebooks (e.g. Jupyter): documenting all steps in the data life cycle
    - Data creation, transformation, graphics etc.
    - Identifying at each stage the relevant participants (PI, researcher, technical staff, tarinees etc.)
    - Long term documentation of the process: http://ssk.huma-num.fr/#/
  - Transient data hosting: metadata, qualification (e.g. future data selection process)
- Dissemination phase
  - Archiving and publishing all intermediary documents (with openness constraints)
  - Final selection and clean up of data sets that should be preserved (with openness constraints)

Cf. Mooc (in English) Recherche reproductible : principes méthodologiques pour une science transparente

The data reuse charter: a tool for agreeing on reuse

- Objective and context
  - Establishing a hand-shake mechanism between the various stakeholders involved in the creation of a data set
    - Researcher, cultural heritage institution, instrument, data hosting infrastructure
  - Work initiated as a collaboration between several major EU organisations
    - DARIAH, CLARIN, E-RIHS, APE, Europeana
  - https://datacharter.hypotheses.org
- Organised as a series of 6 principles
  - Reciprocity, Interoperability, Citability, Openness, Stewardship and Trustworthiness
    - https://datacharter.hypotheses.org/77
- Method
  - Self declaration from each party
    - Can be global to an institution, agreed for a project, specific to an artefact or a data set
  - Availability of questionnaires to guide the design of declarations
- Example
  - DARIAH-Campus

Application: the DARIAH Campus initiative

- Objective: gathering in one single place all DARIAH related training content
  - A variety of objects and sources
    - Videos, texts, training components taken up from previous projects or partners
- Implementing the charter:
  - Global declaration for the platform
  - Citation details for each content
  - Full version available under:
    - https://campus.dariah.eu/docs/dariah-campus-reuse-charter

Work carried out in collaboration with Erzsébet Tóth-Czifra and Toma Tasovac

The DARIAH campus data reuse charter

- Introduction presenting the DARIAH campus vision
  - “Fostering **open access to scholarly resources**, as well as collaboration and fair data-sharing practices among a diverse range of actors involved in knowledge creation in the Arts and Humanities”
  - “as a **mutual declaration of goodwill**, the charter allows us to clarify our expectations regarding the interaction between content creators, users and curators”
- An explicit statement for each of the 6 principles
- An annex (work in progress…) specifying the technical details
  - Metadata profiles, formats and standards used for the various contents

DARIAH campus: data reuse declaration - 1

- Reciprocity
  - “learners ((re)users) are encouraged to share their **feedback** but also **contribute** new training materials”
  - “**contact information** of both authors (together with other contributors) and commenters/reviewers will be made explicit”
- Interoperability
  - “training materials will be made accessible in **open formats**”
  - “share metadata in a **standardized format** to ease harvesting our content”
  - “We make available the content of our hosted learning resources (written in Markdown, a lightweight markup language) available through a **GitHub repository**”
- Citability
  - [We] “make our recommended **citation model** explicit for each training resource shared via DARIAH-Campus”

DARIAH campus: data reuse declaration - 2

- Openness
  - “Open Educational Resource (OER)”
  - “openly and freely available under a **Creative Commons CC-BY 4.0 license**”
- Stewardship
  - “workflows to ensure the **long-term availability** of the training materials hosted on DARIAH-Campus”
- Trustworthiness
  - “we recognize the diverse contributor roles (author, editor, contributor) to clearly document who participated in the production process in order to ensure its **traceability**”

Before we we go further…

- The message once again: open science is just science done right
  - Sharing a culture of sharing
- There are various tools to do so, at different stages of the research process
  - DMPs, Data reuse charter, Laboratory notebooks, SSK, publication repositories
- For those in the TEI landscape, all this seems obvious
  - Let us see why…

The TEI as an open science project

The Text Encoding Initiative for the dummies

- The TEI is an XML application:

  <persName>Tome Geek</persName> went to <placeName>Paris</placeName>.

- The TEI is there for (humanities, but not only) scholars (but also professionals) to encode a variety of cool things:
  - Prose, theatre, poetry, dictionaries (yeah!), manuscripts etc.
- The TEI guidelines provides a vocabulary of 580 elements, a documentation, schemas, tools (e.g. compiler, stylesheets)
- The TEI is a consortium of individual and institutional members
- The TEI is above all a lively community (cf. TEI list)
- It has a thing called ODD, which always sounds mysterious but is mega cool…

The TEI guidelines: an open standard

- What is a standard?
  - The three pillars of a standard: consensus building, public availability, systematic review
  - The TEI consortium is a Standard Development Organisation (SDO)
- What is an open standard?
  - Freely available schemas and documentation (AAMOF all TEI stuff is online)
  - Clear licencing: CC-BY and BSD 2 clauses
- Opening up the standardisation process itself
  - Contribution from the community (GitHub), open discussion process by the TEI technical council
- Relying on a public infrastructure
  - Note: someone has to pay for the service
    - Membership and in-kind contributions
  - Always a tension between service availability and digital independence/sovereignty
    - TEI guidelines development rely on GitHub (acquired by Microsoft)
    - TEI content hosted by Huma-Num (national infrastructure, part of the EU DARIAH infrastructure)

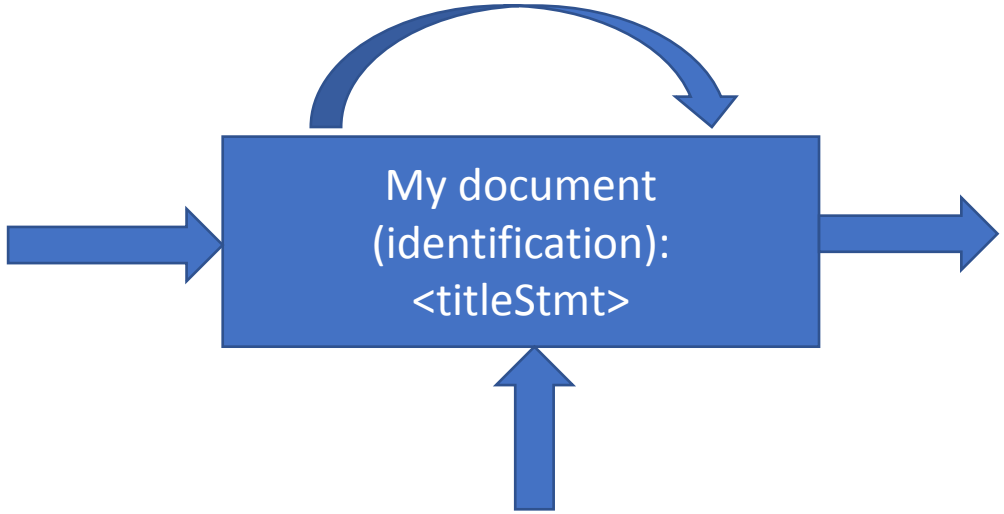
The TEI model as a paragon of well-managed research data

- Never loose track of your documentation: embedded meta-data – **open metadata**
  - The TEI offers a reference framework for documenting a digital document
    - Source – contributions – versions – publication
    - See next slide: the <teiHeader> is complex, but essential for proper digital object management
  - The integration of the header within the document itself is a central concept for information tracing
- Semantic encoding: what you tag is what you mean – **open semantics**
  - The TEI vocabulary is conceived to make the encoding of content as transparent to the underlying semantic as possible
    - Proximity to the scholarly objectives and needs
  - Beyond syntax: the written guidelines (prose, examples) are often more important than the schema
- The next stage: Embedded annotations – **open commentaries**
  - Developing the <standOff> element to provide the same level of requirement on the process of document annotation (human – automatic)
    - Remaining issues: providing a real documentation with the header
- The ODD specification language: documenting choices and practices – **open process**

The teiHeader at a glance

The digital document life cycle (versions):
<revisionDesc>

Origin of the document (physical or digital sources):
<sourceDesc>



Document publication (responsibility, licence, citation):
<publicationStmt>

The production of the document (what, who, how):
<profileDesc>, <respStmt>, <encodingDesc>

Impacting on future practices – fair citing of contributors

- The TEI has introduced a strong culture for documenting contributors to a document
  - Specific elements: <author>, <editor>, <funder>, <principal>, <sponsor>
  - Generic <respStmt> element
- Citing everyone from student to PI
  - Essential for a proper attribution of the work done
  - Richer than any “text” citation (cf. <publicationStmt>)
- An emerging trend
  - Cf. the documentation of roles in a publication: https://casrai.org/credit/
    - Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing

ODD in a nutshell

- ODD: an instance of the Literate programming model (Knuth 1984)
  - One specification comprises
    - The formal definition => from which one can derive, e.g., a RelaxNG schema
    - The prose documentation => html, pdf, docx, ePub etc.
- ODD is a TEI vocabulary => open models
- The TEI guidelines are expressed in ODD
- You can derive TEI customisations from the guidelines using ODD
  - Basis for any project documentation
  - Customisations can be further “chained”
- You can also specify non-TEI vocabularies in ODD
  - E.g. EAD (Romary and Riondet 2018)

Publications can be part of the TEI ecosystem!

- An obvious member of the TEI document landscape (cf. Holmes and Romary 2010)
  - A publication is a text document + stuff: precise metadata, structure, references
  - Customisation facilitates adaptation to specific publication objects (articles, book, slides)
  - Natural continuum with other sources (primary – secondary source continuum)
- Used in various projects and tools
  - Providing a precise snapshot of metadata description: TEI export in HAL (plus bibliographical references acquired by means of Grobid)
  - Publication platforms: Open Edition Journal (JTEI!), Persée.fr, DHQ, Digital mediavalist, DHD conf.
  - Pivot format of the Istex national licence program (https://www.istex.fr, 6 million documents)
  - Management of the “Non patent literature” at the European Patent Office (as part of a 200 million TEI document back office ☺)
- What about JATS?
  - Narrow niche (lack of interoperability with other textual sources), bad standardisation management (sigh…), shaky design (vague TEI plagiarism) - *Vergiss es!*

Consequences for TEI-based projects

- The various ways to be in line with the TEI technical and cultural background:
  - Show the source: always show the scholarly perspective behind a nice looking website
  - Show your ODD: describe the (scholarly) reasons behind the specific data model you've designed
  - Show your software: you never know, it can be useful to one of your followers
  - Going further: showing the process
    - Stages, versions, incomplete digitisation or encoding etc.
  - And public hosting!
  - Beyond citation, post your data reuse charter declaration (in TEI?)

Outline of a possible TEI-centered data reuse charter declaration

- Reciprocity
  - Reciprocal provision of corrections, enrichments, annotations
- Interoperability
  - All content is conformant to the TEI guidelines
- Citability
  - Engagement of citing all sources
  - Provision of a reference citation format in the teiHeader’s
- Openness
  - Provision of the ODD
  - Provision of the source content in TEI
- Stewardship
  - [depending on the project and the available infrastructure]
- Trustworthiness
  - Maintaining content independently of software

TEI Lex 0 as an exemplary open offspring of the TEI

- Making the TEI univocal in a specific subdomain: legacy dictionaries
- From a physical meeting to an immaterial contributors' space
- Adoption by the Elexis project as their pivot format
- Relation to other initiative
  - ISO 24613-4: LMF serialisation in TEI
  - Ontolex: LOD model for lexical data
- Online presence (Kudos to @ttasovac)
- Applying the TEI principles
  - TEI customisation – ODD
  - Open process on GitHub: specification, examples and discussions
  - Mappings to other initiatives: TEI2Ontolex

Point d’orgue – it’s not that easy

- No one knows the TEI
  - Linking with other standardisation bodies with other business models (W3C, ISO)
- Dictionaries and copyright
  - The future is bright, except for the content
- Dealing with the legacy world
  - E.g. Shoebox
- Competition in science
  - Ontolex

TEI for open processes – the SSK

- SSK – Standardisation Survival Kit
  - Designed within the EU Parthenos project
  - On stage: various Parthenos partner, not the least OEAW-ACDH colleagues
- Objective: to describe research processes
  - Initially to pinpoint the role of standards at various stages of such processes
- Concept
  - Description of research scenarios
  - Each scenario is decomposed in various steps
  - Each step points to additional resources (bibliography, tutorial, reference documents)
- Problem to be solve
  - Providing a completely transparent and open description of the underlying content
    - applying standards when describing standards
- Solution
  - The TEI as the description language (<event>!)

Let’s have a ride…

- Mörth Karlheinz, Charles Riondet, Lionel Tadjou, 'Create a dictionary in TEI ', The Standardization Survival Kit (SSK), 2019
  - http://ssk.huma-num.fr/scenarios/SSK_sc_dictionaryInTei

SSK – lessons learnt

- The TEI has been an essential tool in the specification and deployment of the SSK
  - Underlying data modelling possibilities
  - Transparency of descriptions
  - Ease of maintenance (GitHub)
  - Citation made easy
- Towards new realms for the TEI
  - Born digital *performative descriptions*
    - Resembles the concept of didascalies/stage directions
- Past – present – future
  - Hosted by Huma-Num - http://ssk.huma-num.fr
  - Supervised by the DARIAH GIST working group
  - Integration into the SSHOC Market Place

Epilogue – TEI: FAIR by construction

- TEI comes with a culture of open data management
- Data management in a TEI context is centred on the relation to the document
  - Importance of the source, data modelling, dissemination

… but also on the relation between the text and the scholar

  - Encoding as interpretation: the TEI document reflects the research process
- The TEI community has probably been FAIR before the concept existed
  - Hence no need to explain what FAIR is… :-}
- We need to keep advocating such values as part of digital scholarship

Merci pour votre attention !