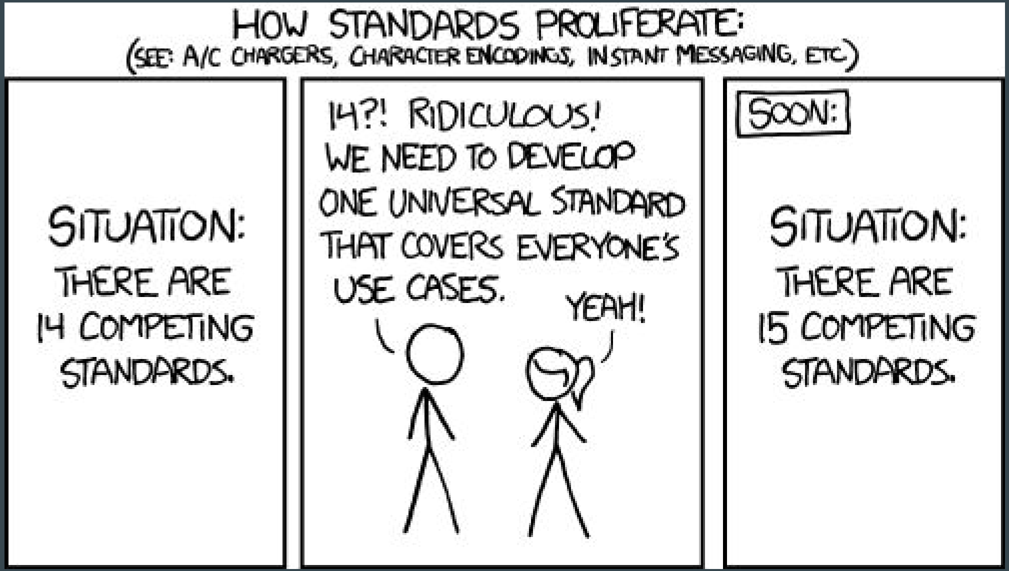
CapiTainS Guidelines

From digital edition to machine actionable edition


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

Thibault Clérice,
PhD at Humboldt Chair of Digital Humanities, Leipzig University
@ponteineptique

A problem (1)

In France alone :

- 1575 PhD for “lettres classiques”
- 1303 for “Archaeology”
- 3579 for “littératures”

- 381 Classics Aggrégation, 48 Grammar, 233 CAPES
- 1682 History Aggrégation, 5347 CAPES
- 1413 French Aggrégation, 3535 CAPES

A problem (2)

- Roughly 12639 MA Thesis in 2015
- 6457 PhD between 2014 and 2016
- 120 Pages / MA Thesis; 400 for PhD

4.099.480 written and lost pages

A problem (3)

"Hasta sub exsertam donec perlata papillam,
haesit uirgineumque alte bibit acta cruo" Aeneid. 11, 803

« Déjà la javeline, pénétrant au dessous de son sein découvert, s'est fixée
immobile : profondément enfoncée, elle a bu son sang virginal. »

P. HEUZE a bien souligné ce que le choix du terme papilla avait de trouble: le terme, à connotation érotique, désigne à proprement parler le bout du sein de la femme et renforce l'image du sang virginal qui s'écoule, faisant du meurtre de Camille un acte proche du viol”

LE CORPS FEMININ ET SES REPRESENTATIONS POETIQUES DANS LA LATINITE TARDIVE,
Sophie MALICK-PRUNIER, 2008, Paris 4, dir. Zarini

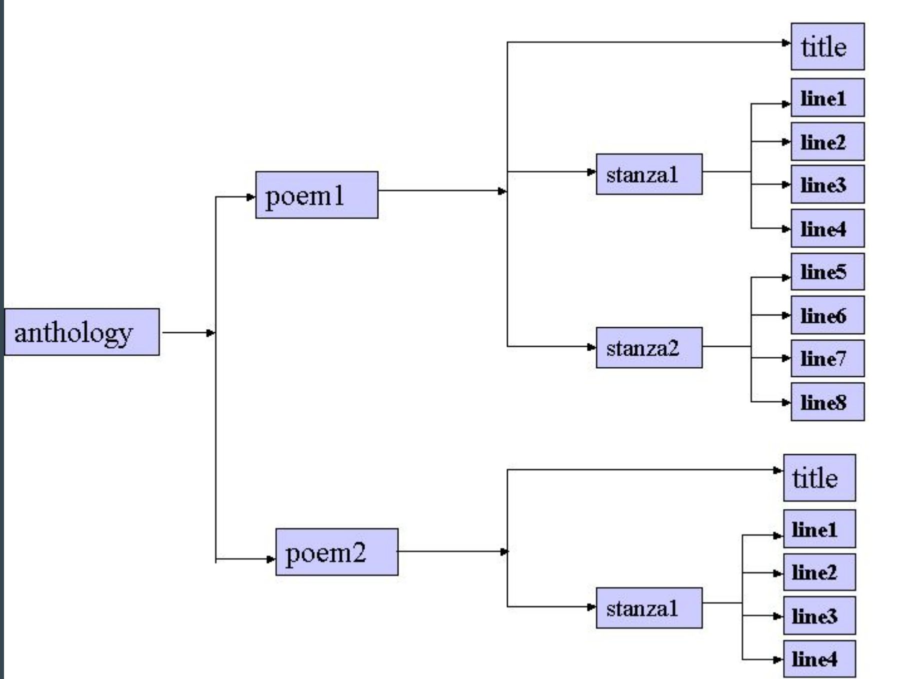
Digital Humanities : OHCO, TEI and CTS (1)

- Digital Humanities is (mainly) about hypertext
- Linking texts and passages
  - Canonical Text Service and Canonical Text Service URN by C. Blackwell and N. Smith
- OHCO : Ordered Hierarchy of Content Object

Digital Humanities : OHCO, TEI and CTS (2)

"Clearly, there are many such trees that might be drawn to describe the structure of this or other anthologies. Some of them might be representable as further subdivisions of this tree: for example, we might subdivide the lines into individual words, since in our simple example no word crosses a line boundary. Surprisingly perhaps, this grossly simplified view of what text is (memorably termed an ordered hierarchy of content objects (OHCO) view of text by Renear et al.20) turns out to be very effective for a large number of purposes."

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html



Digital Humanities : OHCO, TEI and CTS (2)

- Classicists understand Hom. Il. 5.1
  - ἔνθ᾽ αὖ Τυδεΐδῃ Διομήδεϊ Παλλὰς Ἀθήνη
- Classicists even know some of them by heart : Verg. Aen. 1.1
  - Arma virumque cano, Troiae qui primus ab oris
- But nobody else actually knows what Cic. De fin. 1.5 means
  - Leaked password ?
  - A French Bank institution ?

Digital Humanities : CTS (1)

The **Canonical Text Services** Protocol is a specification that "defines a network service for identifying texts and for retrieving fragments of texts by canonical reference expressed as CTS-URNs." CTS and CTS-URN provides an interoperable, open and persistent system for sharing text resources and parts of them on the web.

At the core of the CTS URN is the idea of representing texts as part of a graph, where nodes resolve to texts, objects or images, and the edges provide navigation between them. Text nodes themselves consist of citable nodes, with each node having the following properties:

- belongs to a specific version of a work in a FRBR-like hierarchy
- belongs to a citation hierarchy of 1 or more levels
- is ordered
- may have mixed content

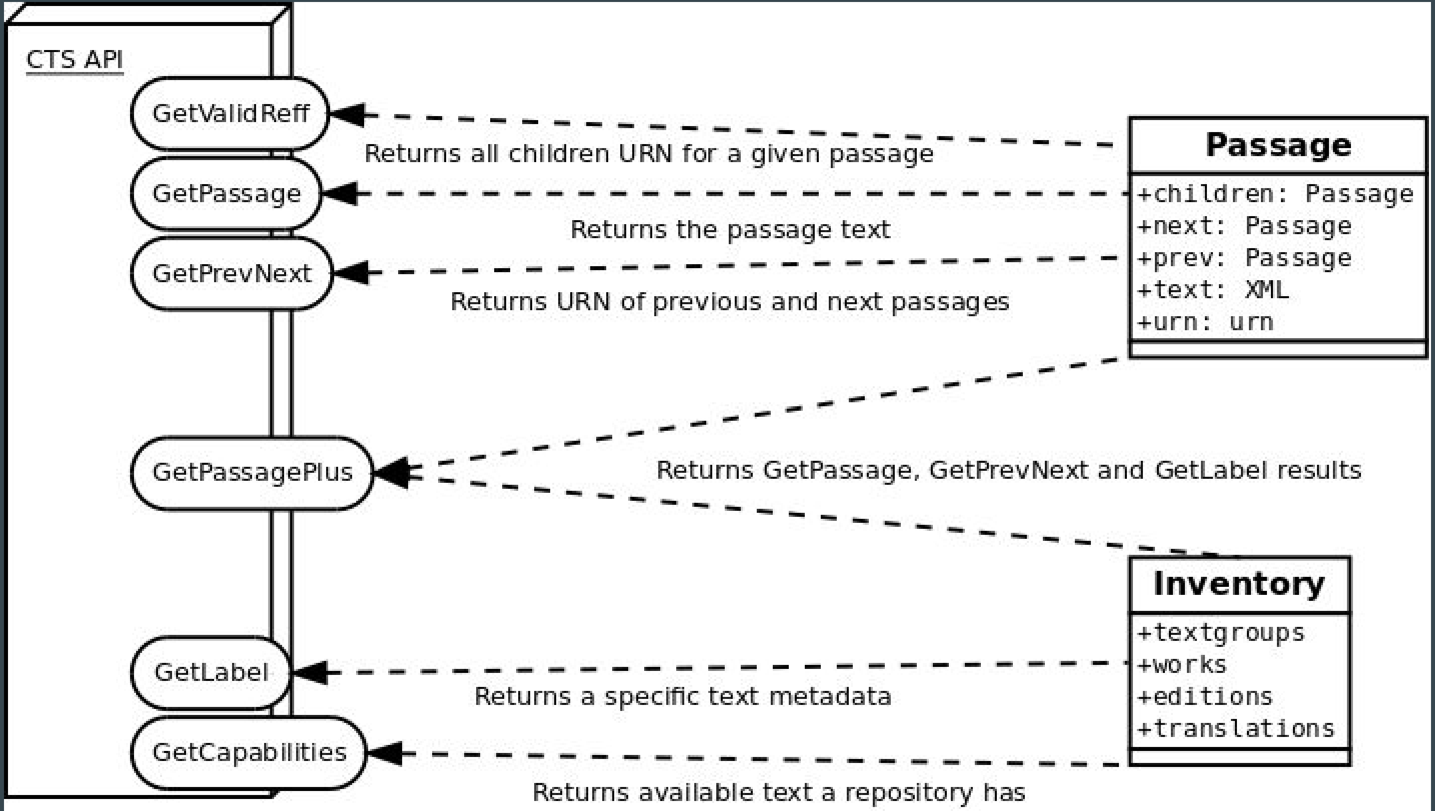
Neel Smith and Chris Blackwell, Canonical Text Services , http://cite-architecture.github.io

Digital Humanities : CTS (2)

urn:	cts:	greekLit:	tlg0012.	tlg001.	perseus-grc1:	1.	1
URN namespace		*CTS namespace*	*Textgroup eg Author*	*Work Identifier*	*Version Identifier*	*Reference*	*Subreference*
		Ancient Greek Literature	Homeric Texts	Illiad	First version edited on Perseus	Book 1	Line 1

urn:cts:latinLit:phi1294.phi002.perseus-lat2 -> Martial, Epigrammata
urn:cts:froLit:jns915.jns1856.ciham-fro1 -> Wauchier de Denain, Vie de Saint Martin
urn:cts:pdlpsci:bodin.livrep.perseus-fre1 -> Bodin, Six Books of a Commonweale
urn:cts:latinLit:phi0690.phi003.perseus-lat1:1.1 -> (Virgile,Virg. Uirg. Verg...), (Aeneid, Énéide, Éné.) 1.1

Digital Humanities : CTS (3)



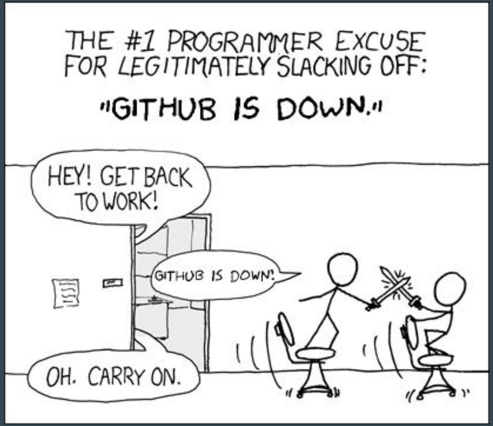
CapiTainS : CTS API and ain ?

http://capitains.github.io - http://github.com/Capitains

Bridget Almas (Perseids), Frederik Baumgardt (Perseids), Thibault Clérice (Humboldt Chair of DH)

CapiTainS : Why Open Source ?

- Lack of funding
  - More people to work on your software
  - What happens after the project funds run out ?
- Visibility effort
  - More people to work with your software
  - Organization, Project and Funder should be known so that you continue getting funded
- Ability to work in team across offices and borders for **free**
  - **Thanks GITHUB, Travis and others !**
- Transparency issue
  - People can check what you stated
- Sharing.
- Lot of other reasons.



CapiTainS : How to do Open Source ?

- **Documentation**, because there is no way a stranger will spend hours reading your code
- **Good documentation**, because bad documentation is still documentation
- **Setup documentation** is also documentation
- **Automated tests**, because I want to know if someone breaks my code, including me
- **Good versioning**
- **Public communication** : issues and pull requests are not only fancy tools on github,
- Did I say **documentation** ?

COMMENT	DATE
CREATED MAIN LOOP & TIMING CONTROL	14 HOURS AGO
ENABLED CONFIG FILE PARSING	9 HOURS AGO
MISC BUGFIXES	5 HOURS AGO
CODE ADDITIONS/EDITS	4 HOURS AGO
MORE CODE	4 HOURS AGO
HERE HAVE CODE	4 HOURS AGO
AAAAAAAA	3 HOURS AGO
ADKFJSLKDFJSDKLFJ	3 HOURS AGO
MY HANDS ARE TYPING WORDS	2 HOURS AGO
HAAAAAAAANDS	2 HOURS AGO

AS A PROJECT DRAGS ON, MY GIT COMMIT MESSAGES GET LESS AND LESS INFORMATIVE.

CapiTainS : Code Base

Sparrow

My Capytain

Cavern

Ahab

Nemo

CTS API

Inventory Maker

Hook Test

Hook

http://capitains.github.io/pages/tools.html

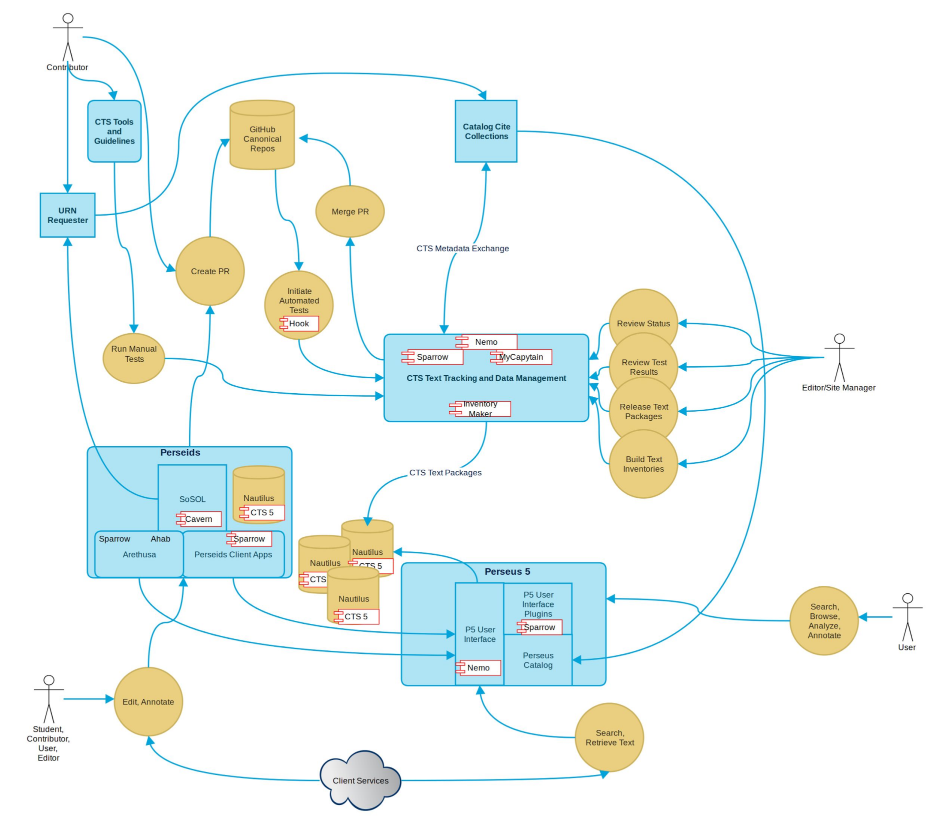
CapiTainS : Code Base (2)

http://www.gliffy.com/go/publish/7879353

http://cts.perseids.org

http://ci.perseids.org

http://www.perseids.org/sites/joth/



CapiTainS : Guidelines (1)

- Explicitly state CTS related information in files
- CTS is
    - Texts
    - Metadata
- Divide responsibilities:
    - Browsing catalog
    - Browsing text content

http://capitains.github.io/pages/guidelines.html

CapiTainS Guidelines (2) : Directory structure

```
data/
    |- textgroup
        |-__cts__.xml
        |- work
            |-__cts__.xml
            |- part-of-the-urn.xml (phi1294.phi001.perseus-lat2.xml)
```

CapiTainS Guidelines (3) : TEI File (Epidoc version)

TEI/teiHeader/encodingDesc

```
<refsDecl n="CTS">
 <cRefPattern n="line" matchPattern="(.+).(.+)"
replacementPattern="#xpath(/tei:TEI/tei:text/tei:body/tei:div[@n=’$1’]//tei:l[@n=’$2’])">
 <p>This pointer pattern extracts book and line</p>
 </cRefPattern>
 <cRefPattern n="book" matchPattern="(.+)"
replacementPattern="#xpath(/tei:TEI/tei:text/tei:body/tei:div[@n=’$1’])">
 <p>This pointer pattern extracts book.</p>
 </cRefPattern>
</refsDecl>
```

TEI/text/body

```
<div type=”edition|translation” n=”urn:cts:latinLit:phi1294.phi002.perseus-lat2”>
…
</div>
```

CapiTainS Guidelines (4) : Metadata files

Group level

```
<ti:textgroup xmlns:ti="http://chs.harvard.edu/xmlns/cts" urn="urn:cts:latinLit:phi1294">
 <ti:groupname xml:lang="eng">Martial</ti:groupname>
</ti:textgroup>
```

Work level

```
<ti:work xmlns:ti="http://chs.harvard.edu/xmlns/cts" groupUrn="urn:cts:latinLit:phi1294"
urn="urn:cts:latinLit:phi1294.phi002">
 <ti:title xml:lang="eng">Epigrammata</ti:title>
 <!-- For each "text", either edition or translation, there should be a ti:edition or ti:translation node -->
 <ti:edition workUrn="urn:cts:latinLit:phi1294.phi002" urn="urn:cts:latinLit:phi1294.phi002.perseus-lat2">
 <ti:label xml:lang="eng">Epigrammata</ti:label>
 <ti:description xml:lang="eng">
 M. Valerii Martialis Epigrammaton libri / recognovit W. Heraeus
 </ti:description>
 </ti:edition>
</ti:work>
```

CapiTainS : Open Data Easy, Linked Data Easy (1)

- As a Data Provider
  - Run your own website locally in less than 10 minutes (https://youtu.be/_Vmwz_761GM)
  - Make API following standards easily
  - Profit (and participate) in a coding comunity ?
- As a Data Consumer
  - Create interfaces using CTS data easily
  - Parse texts locally for Natural Language Processing
  - Parse texts from different APIs

ONLINE
AVAILABILITY
RE-USE RIGHTS
...so people can do interesting things in new ways with the materials.
Nick Shockey and Jonathan Eisen
WWW.PHDCOMICS.COM

https://www.youtube.com/watch?v=L5rVH1KGBCY

CapiTainS : Open Data Easy, Linked Data Easy (2)

Current Data Providers :

- PerseusDL Latin (676 Files), Ancient Greek (1429), Norse, English, Secondary Sources (Dictionaries, Lexicons...)
- Perseids new texts
- OpenGreekAndLatin : CSEL (278), First1KGreek (311), Patrologia Latina
- Persian Digital Library (PersDigUMD)
- Incoming PhD on Medieval French at CIHAM, Lyon
- Incoming Hyperdonat project
- Incoming OpenArabic

CapiTainS : Open Data Easy, Linked Data Easy (3)

- Lowered skill threshold to create CTS resources
- Open contribution
- Next step ?
  - Perseids Hackathon (May 9th - 14th)
  - OAI-PMH Layer
  - New Inventory Maker for Python
  - More documentation
  - More training (DH2016; Lyon HiSoMA Lab; CHS)
  - PhP Abstraction is welcome if you do PhP
  - Java Abstraction is as welcome if you do Java (And I want *CapiTainS Javara* to become a thing)

Thanks

Special thanks to the team at Perseids and DH Chair as well as R. Lopes for the logos !

Useful links :

- http://capitains.github.io
- capitains@googlegroups.com
- http://github.com/capitains

CapiTainS

See you at the training !

Contact : thibault.clerice@uni-leipzig.de / @ponteineptique (GitHub and twitter)