USING DATATAGS TO CLASSIFY PERSONAL DATA UNDER GDPR

Rob Baxter, EPCC

Emily Thomas, Heiko Tjalsma, DANS

The University of Edinburgh: Dealing with Data 2017

Data Archiving and Networked Services
DANS

EUDAT

epcc

Background: EPCC & EUDAT

- **EPCC:**
  - the high-performance computing & data centre here at the University
  - 90 staff, externally funded, outward facing, project based
  - host major UK computing systems & engage in European projects including…
- **EUDAT** (FP7, 2011-14) and **EUDAT2020** (H2020, 2014-18)**:**
  - 30+ European partners: HPC centres, data repositories, research infrastructures
  - developing common approaches to research data management in…
- the Collaborative Data Infrastructure (**CDI**):
  - a federation of repositories and services providers connected at several levels:
    - technical and service infrastructure
    - policy and best practice
    - community working groups and training
  - creating the data foundation for…
- the European Open Science Cloud (**EOSC**)
  - the federated future of European research IT systems (?)

Data Archiving and Networked Services DANS

epcc

EU General Data Protection Regulation (GDPR)

- Passed 14 April 2016, enforceable from 25 May 2018 (!)
- A European *Regulation*, not a European Directive
  - although Data Protection Authorities remain national
  - derogations possible for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes
  - and Codes of Conduct can be national or drafted by “learned bodies”
- Enshrines new rights for data subjects
  - Article 15: Right of access by the data subject
  - Article 16: Right to rectification
  - Article 17: Right to erasure (to be forgotten)
  - Article 18: Right to restriction of processing
  - Article 19: Notification obligation regarding rectification
  - Article 20: Right to data portability
  - Article 21: Right to object
- Informed consent as basis for use of personal data
- Requires data minimisation & “privacy by design” in DM services
- Extremely high fines for trespassing (data leakage!)

The DataTags model

- Sweeney, Crosas & Bar-Sinai (Harvard 2015): a *DataTags repository*
  - *Sharing Sensitive Data with Confidence: The Datatags System*
    - Technology Science [Internet], 2015. http://techscience.org/a/2015101601/
- Stores and shares data objects in accordance with different security levels, access requirements and usage agreements, encoded as a *data tag*
  - based on American laws and legislations of personal data
- Can we apply DataTags to GDPR?

Tag Type	Description	Security Features	Access Credentials
Blue	Public	Clear storage, Clear transmit	Open
Green	Controlled public	Clear storage, Clear transmit	Email- or OAuth Verified Registration
Yellow	Accountable	Clear storage, Encrypted transmit	Password, Registered, Approval, Click-through DUA
Orange	More accountable	Encrypted storage, Encrypted transmit	Password, Registered, Approval, Signed DUA
Red	Fully accountable	Encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA
Crimson	Maximally restricted	Multi-encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA

Adapting DataTags to GDPR: DANS pilot project

- EUDAT goal: *"define categories of data sensitivity compatible with national and European regulations; and develop consistent guidelines for restricted data access to be adopted in the CDI"*

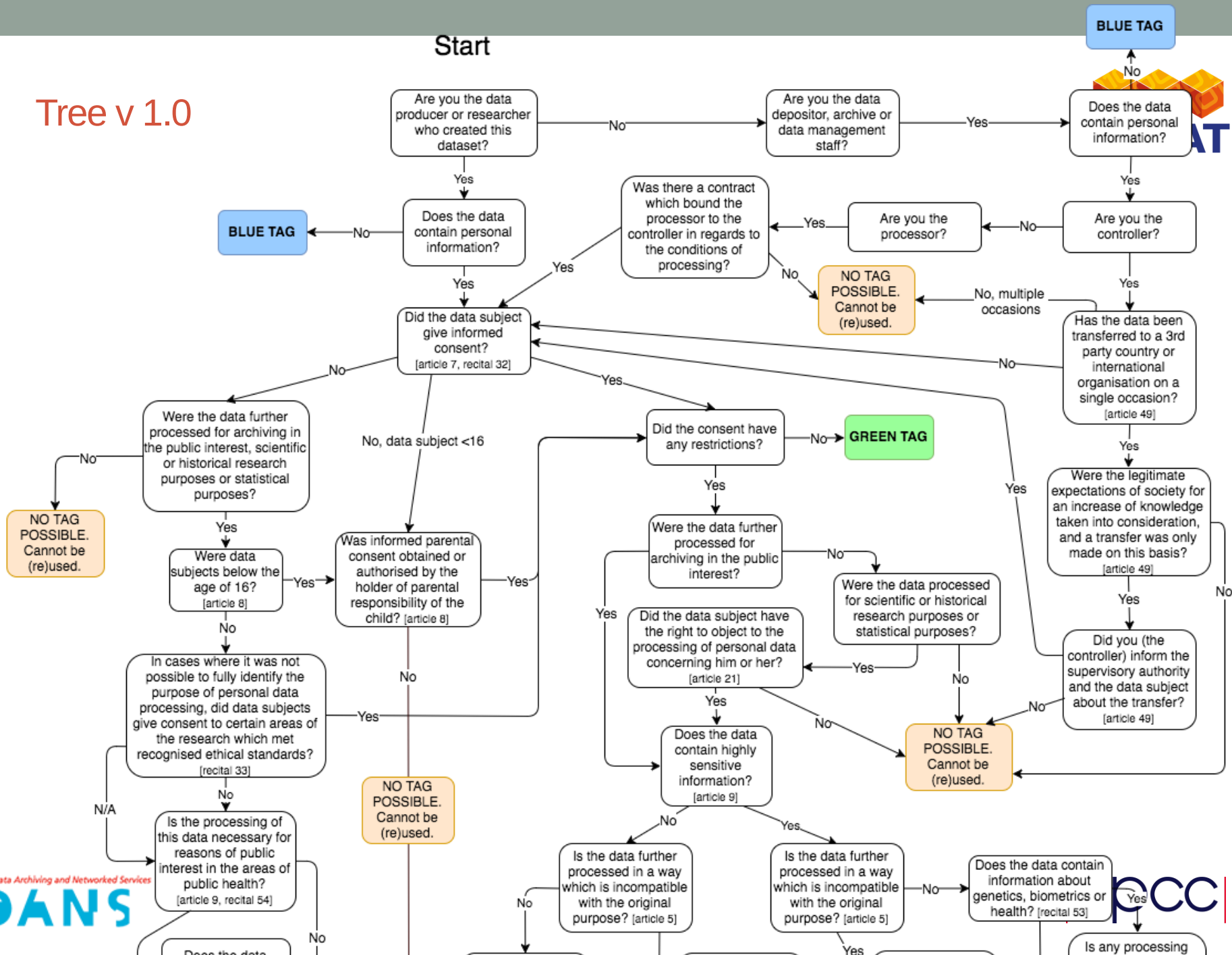
1. Identify the relevant articles of the GDPR for research and archive purposes
   - example: Article 9(2) sets out the circumstances in which the processing of sensitive personal data (which is otherwise prohibited) may take place:
     - "necessary for archiving purposes in the public interest, or scientific and historical research purposes or statistical purposes in accordance with Article 89(1)"
2. Transform relevant Articles into questions
   - were the data processed for archiving in the public interest, scientific or historical research purposes or statistical purposes?
   - would you consider the dataset to contain sensitive personal information?

Adapting DataTags to GDPR

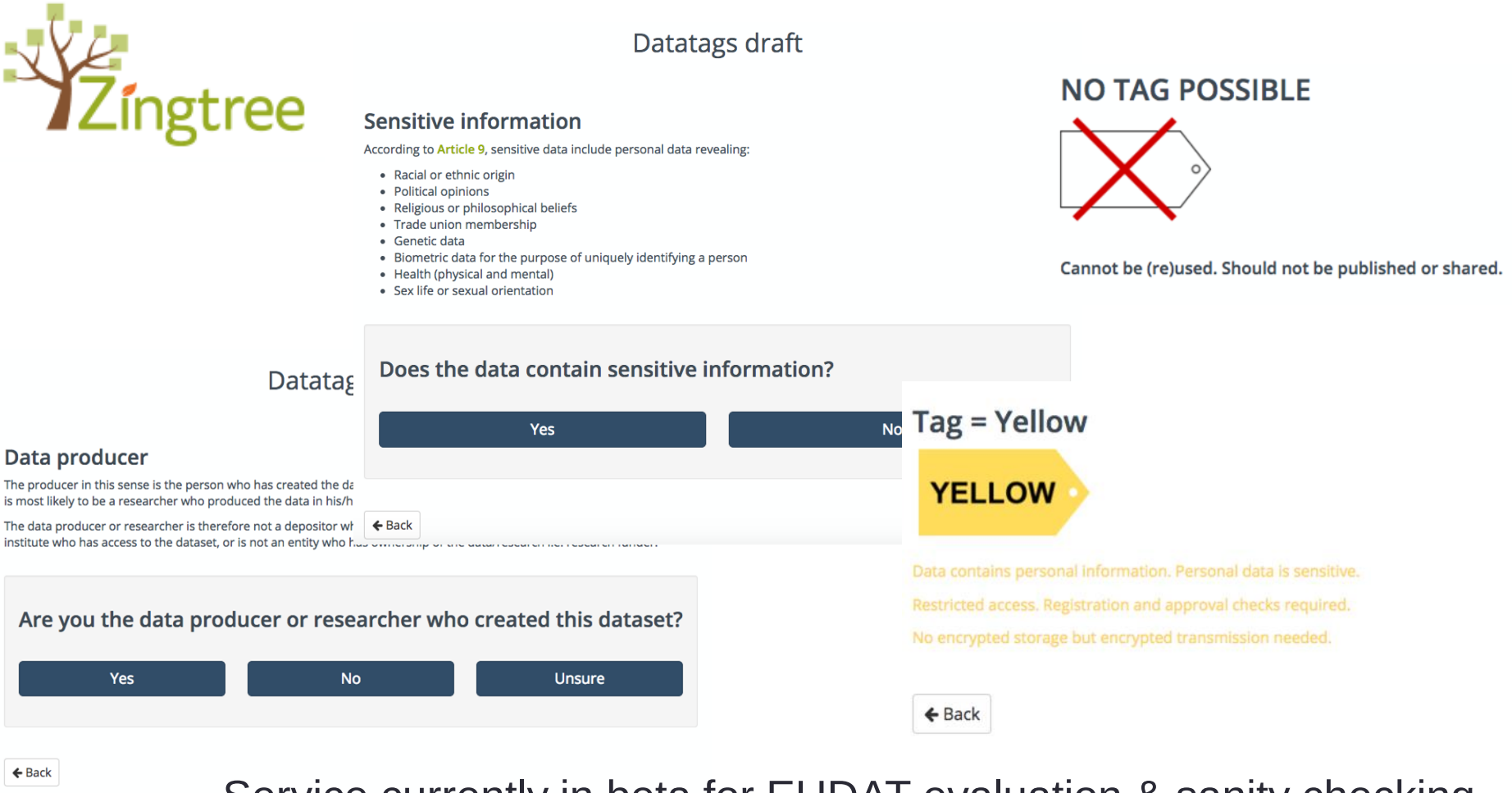
3. Evolve into a decision tree
   - create routes for questions, ending with tags
   - decide on tag options and recommendations following each route
     - settled on 4 tags:

Tag type:	Authentication	When transmitted	When stored	Reading/downloading rights
0. Public access (non-personal data)	None needed	Without encryption, with checksum	Standard - clear storage	Everyone (with or without registration)
1. Basic access (non-confidential personal data)	Registration necessary	Without encryption, with checksum	Standard - clear storage	All registered users
2. Restricted access (sensitive personal data)	Registration via repository and approval of depositor	With encryption, with checksum	Standard - clear storage	All registered users, *after approval of depositor*
3. Selected access (highly sensitive data)	Registration via repository and mandatory further identification	Multi-encryption, with checksum	Not accessible via the internet and with encryption	NOT via repository, checked users only

Tree v 1.0


Start
Are you the data producer or researcher who created this dataset?
No
Are you the data depositor, archive or data management staff?
Yes
Does the data contain personal information?
No
BLUE TAG
Yes
Does the data contain personal information?
No
BLUE TAG
Yes
Was there a contract which bound the processor to the controller in regards to the conditions of processing?
Yes
Are you the processor?
No
Are you the controller?
No
NO TAG POSSIBLE. Cannot be (re)used.
No, multiple occasions
Yes
Did the data subject give informed consent? [article 7, recital 32]
No
Has the data been transferred to a 3rd party country or international organisation on a single occasion? [article 49]
No
Yes
Were the data further processed for archiving in the public interest, scientific or historical research purposes or statistical purposes?
No
NO TAG POSSIBLE. Cannot be (re)used.
Yes
Were data subjects below the age of 16? [article 8]
Yes
Was informed parental consent obtained or authorised by the holder of parental responsibility of the child? [article 8]
No, data subject <16
Yes
No
In cases where it was not possible to fully identify the purpose of personal data processing, did data subjects give consent to certain areas of the research which met recognised ethical standards? [recital 33]
Yes
No
NO TAG POSSIBLE. Cannot be (re)used.
N/A
Is the processing of this data necessary for reasons of public interest in the areas of public health? [article 9, recital 54]
No
Did the consent have any restrictions?
No
GREEN TAG
Yes
Were the data further processed for archiving in the public interest?
No
Were the data processed for scientific or historical research purposes or statistical purposes?
Yes
Did the data subject have the right to object to the processing of personal data concerning him or her? [article 21]
Yes
No
Yes
No
Does the data contain highly sensitive information? [article 9]
No
Yes
Is the data further processed in a way which is incompatible with the original purpose? [article 5]
No
Is the data further processed in a way which is incompatible with the original purpose? [article 5]
No
Yes
Does the data contain information about genetics, biometrics or health? [recital 53]
Yes
Is any processing
Were the legitimate expectations of society for an increase of knowledge taken into consideration, and a transfer was only made on this basis? [article 49]
Yes
No
Did you (the controller) inform the supervisory authority and the data subject about the transfer? [article 49]
No
NO TAG POSSIBLE. Cannot be (re)used.

4. Zingtree (www.zingtree.com)



- Service currently in beta for EUDAT evaluation & sanity checking
  - https://zingtree.com/host.php?tree_id=442670046

Implementing data tags across federated repositories

- A data object will have one (and only one) data tag
- Where to record data tags?
  - in the data objects' Handle records (or DOIs or…)
    - globally visible
  - in service-local metadata databases
    - at "entry point" of data into the CDI infrastructure (the "repository of record")
    - needs to propagate into replicas of data under control of remote DM services
  - in global catalogue records
    - as part of the standard OAI-PMH metadata publish/subscribe
    - but the catalogue service provider is not necessarily the data processor who needs to know the sensitivity tag!
- One, some or all?
  - keeping sync'ed will be an issue, as will be…

DataTags issues

- Granularity
  - at what level does one tag a “data object”?
- Binding
  - how to maintain correspondence between tag & object in a tamper-resistant way?
- Encryption
  - who holds the keys?
- Time & events
  - tags might change with time (e.g. children grow up, subjects die)
  - propagating changes across distant replicas is non-trivial
- Data or metadata?
  - one person’s metadata is another’s data

Conclusions

- GDPR across a multi-organisational distributed infrastructure is going to be a challenge (!)
- Data admins/repositarians need infrastructure support
- DataTags could be a win for automatic, rule-based management
  - systems like iRODS could read tags & trigger actions
  - adding to Handle records would be favourite
- Understanding how they can change with time, events needs thought!
- Further work could look at using the same approach for codes of conduct, ethical frameworks…

Acknowledgements

- At DANS
  - Peter Doorn, Ingrid Dillo & Heiko Tjalsma for the idea
  - Emily Thomas for doing all the work
  - colleagues at the workshop on 30 August
- At CERN
  - David Foster, Head of Data Privacy Protection, for sanity checking
- At EUDAT
  - Valentino Cavalli, Helen Frew, Melanie Imming, Catherine Inglis, Francesca Iozzi, Simon Lambert, Damien Lecarpentier, Simone Sacchi