USING DATATAGS TO CLASSIFY PERSONAL DATA UNDER GDPR

Rob Baxter, EPCC
Emily Thomas, Heiko Tjalsma, DANS

The University of Edinburgh: Dealing with Data 2017
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 (?)
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*

- 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?

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

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

- Service currently in beta for EUDAT evaluation & sanity checking
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