<table>
<thead>
<tr>
<th><strong>Document type:</strong></th>
<th>Other document (Open)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title:</strong></td>
<td>WG2N1953 KG input re Datasets and Data Provenance</td>
</tr>
<tr>
<td><strong>Status:</strong></td>
<td>Koren contribution for Dataset and Data Provenance from Beijing Plenary 2014</td>
</tr>
<tr>
<td><strong>Date of document:</strong></td>
<td>2014-09-17</td>
</tr>
<tr>
<td><strong>Source:</strong></td>
<td>Koren Delegation</td>
</tr>
<tr>
<td><strong>Expected action:</strong></td>
<td>INFO</td>
</tr>
<tr>
<td><strong>Email of secretary:</strong></td>
<td><a href="mailto:timothy@schoechle.org">timothy@schoechle.org</a></td>
</tr>
<tr>
<td><strong>Committee URL:</strong></td>
<td><a href="http://isotc.iso.org/livelink/livelink/open/jtc1sc32wg2">http://isotc.iso.org/livelink/livelink/open/jtc1sc32wg2</a></td>
</tr>
</tbody>
</table>
Contribution on the specification for the registration of
datasets and the provenance of data

*(Individual expert contribution)*

At the Beijing SC32 Plenary (June 2014) resolutions were passed to allow WG2 to enter into two
news areas of work, the registration of datasets and the registration of the provenance of data.

I recall that two years ago I developed a specification for a Data Catalogue, the purpose of that
catalogue being the recording of the “nature, validity, ownership and whereabouts” of the datasets
available to the client organisation.

I attach an extract from that specification as an initial contribution to our deliberations on these
matters.

*Keith Gordon*
*14 June 2014*
1. **Purpose:**

1.1. The purpose of the Data Catalogue is to provide a means for identifying what data is available to users within XXXX. It is to contain information on the nature, validity, ownership and whereabouts of data.

1.2. There is to be only one Data Catalogue within XXXX.

1.3. This is to meet the general requirement that, to ensure the consistent and appropriate use of data and information, and to prevent duplication of work, the existence of data and information and their context must be easily discoverable by XXXX staff and contractors. This approach is consistent with those in place to respond to Freedom of Information and Environmental Information Requests, and XXXX’s Transparency Policy.

2. **Client:** The client is the XXXX Data Assurance Manager, who is tasked with providing assurance that data is being appropriately used within the technical work programme and monitoring the application of the data management process.

3. **Users:** There are two classes of users:

3.1. **Data Catalogue Steward:** This is a role that will be tasked with the maintenance of the information within the catalogue and will, therefore require create, read, update and delete access to all information within the catalogue. There may be more than one person with this role, but the number of users with this role will be restricted.

3.2. **General User:** This may be any member of the XXXX staff or any contractor employed by XXXX.

4. **Definitions:** The following definitions are applicable for this specification:

4.1. **Accuracy:** A measure of the quality of data that represents the difference between an estimate of a parameter and its true value.

4.2. **Completeness:** A measure of the quality of data that represents the degree to which data and information address the entire spatial or temporal scope of the domain and are of sufficient precision.

4.3. **Consistency:** A measure of the quality of data that represents the use of standards and consistent terminology and methods.

4.4. **Data:** [Definition to follow – use the definition in the revised policy document]

4.5. **Data owner:** The XXXX employee who has accountability for the appropriate use of data and information - its security, integrity, quality and availability - and for ensuring that data and information management procedures are followed; the owner may delegate the day-to-day responsibility to one or more stewards.

4.6. **Data set:** A collection of data usually presented in a tabular form; these may be stored in tables in word processed documents, spread sheets or databases.

4.7. **Data set status:** One of five possible statuses of a data set, viz:

4.7.1. **Under procurement:** XXXX are in the process of procuring the data set.
4.7.2. **Live**: The data set is a current data set in use within XXXX.

4.7.3. **Superseded**: The data set has been replaced by an updated data set.

4.7.4. **Retired**: The data set is still in existence within XXXX but is no longer being used.

4.7.5. **Disposed of**: The data set is no longer in existence within XXXX.

4.8. **Data steward**: The XXXX employee who has responsibility for the management of data or information assets and making adequate provision for the long-term care of data and information.

4.9. **Descriptor**: An additional privacy marking that is added to protective marking which identifies sensitivities around distribution and handling; examples of include “Budget”, “Commercial”, “Management” and “Personal”.

4.10. **Information Asset**: A body of information, defined and managed as a single unit so it can be understood, shared, protected and exploited effectively. [Note: this definition is from the YYYY, how does it differ from the definition of dataset?]

4.11. **Keyword**: A word used to describe the purpose and/or coverage of a data set.

4.12. **Nationality caveat**: An additional marking that is added to protective marking defining which nationality groups information may be released to, such as “UK EYES ONLY”.

4.13. **Relevance**: A measure of the quality of data that represents the degree to which data and information meet users’ needs.

4.14. **Security classification**: One of five levels of protective marking (PROTECT, RESTRICTED, CONFIDENTIAL, SECRET and TOP SECRET) applied to information to ensure that it is correctly handled; information without a classification may be marked UNCLASSIFIED.

4.15. **Transparency**: A measure of the quality of data that represents the provision of documentation about the provenance, assumptions, methods, and limitations of data and information.

5. **Assumptions**: The following assumptions have been used in drawing up this specification:

5.1. Data sets may be pre-registered before they are procured.

5.2. Data sets may reside in a number of different locations.

5.3. Data sets may exist in a number of different storage formats (on paper, in spreadsheets, in databases, etc).

5.4. Each data set may have multiple copies.

5.5. The only data sets within scope are those data sets that contain structured data that is relevant to the development of safety cases.

5.6. Assessments of the quality of the data within a data set will only be made at the data set level.

5.7. Assessments of the quality of the data within a data set will require assessments of accuracy, relevance, transparency, consistency and completeness. How these assessments are to be specified is to be the subject of further study.

5.8. The identifier of an XXXX employee will identify the post and department of that employee.

5.9. There may be ‘competing’ data quality assessments.
5.10. Where data sets are stored in databases, these databases are managed with an SQL-based database management system, for example, Oracle, Microsoft SQL Server, Microsoft Access.

5.11. Updating (creation, amendment or deletion of records) of information in the catalogue is to be limited to a small number of individuals (or roles) but read access is to be more widely available, preferably over the XXXX intranet.

6. Data requirements: For each data set the information shown in the conceptual model at Page 8 and described below is to be able to be recorded:

6.1. A unique identifier of the data set.

6.2. The name of the data set.

6.3. The version of the data set.

6.4. A description of the purpose of the data set.

6.5. The unique identifier of the Information Asset of which this data set is, or will be, a part.

6.6. The domain (or business area) covered by the data set; this will normally be represented by a set of keywords.

6.7. The full security classification of the data set, including:

6.7.1. The security classification itself.

6.7.2. Any descriptor ('Budget', 'Commercial', etc) attached to the security classification.

6.7.3. Any nationality caveat ('UK EYES ONLY', etc) attached to the security classification.

6.8. If appropriate, a statement that the data set is in the process of being procured.

6.9. If appropriate, a statement that the data set is 'live', with the effective date of this status.

6.10. If appropriate, a statement that the data set is superseded (ie has been replaced by an updated data set), with the effective date of this status.

6.11. If appropriate, a statement that the data set is retired (ie, still in existence within XXXX but not being used), with the effective date of this status.

6.12. If appropriate, a statement that the data set has been disposed of, with a statement of the method of disposal and the date of disposal.

6.13. If the data set already exists (ie, is 'live', is superseded or is retired), the primary storage location of the data set.

6.14. If the data set already exists, the storage format of the data set at the primary location.

6.15. If the data set already exists and any additional copies of the data set are held, for each copy:

6.15.1. The storage location of that copy.

6.15.2. The storage format of that copy.

6.16. The data owner of the data set.

6.17. The nominated data steward for the data set.
6.18. If known, the source of the data set.

6.19. If known, the origination date of the data set.

6.20. The date that XXXX first procured the data set.

6.21. If the data set is superseded:
   6.21.2. The effective date of the supersession.
   6.21.3. The reason for the supersession.

6.22. If any, assessments of the quality of the accuracy of the data within the data set including:
   6.22.1. A statement of the assessment of the accuracy of the data within the data set.
   [Note: how will this is to be specified is to follow]
   6.22.2. A statement of the confidence level of the assessment of the accuracy of the data within the data set.
   6.22.3. A statement of the rationale for that assessment of the accuracy of the data within the data set.
   6.22.4. The identifier of the person making the assessment of the accuracy of the data within the data set.
   6.22.5. The effective date of the assessment of the accuracy of the data within the data set.

6.23. If any, assessments of the quality of the relevance of the data within the data set including:
   6.23.1. A statement of the assessment of the relevance of the data within the data set.
   [Note: how will this is to be specified is to follow]
   6.23.2. A statement of the confidence level of the assessment of the relevance of the data within the data set.
   6.23.3. A statement of the rationale for that assessment of the relevance of the data within the data set.
   6.23.4. The identifier of the person making the assessment of the relevance of the data within the data set.
   6.23.5. The effective date of the assessment of the relevance of the data within the data set.

6.24. If any, assessments of the quality of the transparency of the data within the data set including:
   6.24.1. A statement of the assessment of the provenance, assumptions, methods and limitations of the data within the data set.
   [Note: how will this is to be specified is to follow]
   6.24.2. A statement of the confidence level of the assessment of the provenance, assumptions, methods and limitations of the data within the data set.
   6.24.3. A statement of the rationale for that assessment of the provenance, assumptions, methods and limitations of the data within the data set.
   6.24.4. The identifier of the person making the assessment of the provenance, assumptions, methods and limitations of the data within the data set.
6.24.5. The effective date of the assessment of the provenance, assumptions, methods and limitations of the data within the data set.

6.25. If any, assessments of the quality of the consistency of the data within the data set including:

6.25.1. A statement of the assessment of the use of standards and consistent terminology and methods in development of the data within the data set. [Note: how will this is to be specified is to follow]

6.25.2. A statement of the confidence level of the assessment of the use of standards and consistent terminology and methods in development of the data within the data set.

6.25.3. A statement of the rationale for that assessment of the use of standards and consistent terminology and methods in development of the data within the data set.

6.25.4. The identifier of the person making the assessment of the use of standards and consistent terminology and methods in development of the data within the data set.

6.25.5. The effective date of the assessment of the use of standards and consistent terminology and methods in development of the data within the data set.

6.26. If any, assessments of the quality of the completeness of the data within the data set including:

6.26.1. A statement of the assessment of the degree to which the data within the data set address the entire spatial or temporal scope of the domain and are of sufficient precision. [Note: how will this is to be specified is to follow]

6.26.2. A statement of the confidence level of the assessment of the degree to which the data within the data set address the entire spatial or temporal scope of the domain and are of sufficient precision.

6.26.3. A statement of the rationale for that assessment of the degree to which the data within the data set address the entire spatial or temporal scope of the domain and are of sufficient precision.

6.26.4. The identifier of the person making the assessment of the degree to which the data within the data set address the entire spatial or temporal scope of the domain and are of sufficient precision.

6.26.5. The effective date of the assessment of the degree to which the data within the data set address the entire spatial or temporal scope of the domain and are of sufficient precision.

6.27. If any, details of appropriate XXXX projects for the use of the data set, including:

6.27.1. The name of the project.

6.27.2. Whether the data set was used by the project.

6.27.3. If the data set was not used, the reason for the non-usage.

6.28. The date the data set is due for review to re-assess the quality and the security of the data set or whether the data set is due for retirement.

6.29. Any other comment on the data set not covered by the above.
7. **Functional Requirements**: The system shall:

7.1. Allow a user designated as a Data Catalogue Steward to create a catalogue entry for a data set.

7.2. Allow a user designated as a Data Catalogue Steward to retrieve an existing catalogue entry for a data set from a 'search list' that is generated using any one of the following search criteria:

   7.2.1. The unique identifier of the data set.
   7.2.2. The name of the data set.
   7.2.3. A search string that is included in the description of the purpose of the data set.
   7.2.4. A search string that includes some or all of the keywords that represents the domain (or business area) covered by the data set.

7.3. Allow a user designated as a Data Catalogue Steward to read and print a retrieved catalogue entry for a data set.

7.4. Allow a user designated as a Data Catalogue Steward to update a retrieved catalogue entry for a data set.

7.5. Allow a user designated as a Data Catalogue Steward to delete a retrieved catalogue entry for a data set.

7.6. Allow a user designated as a General User to retrieve, read and print an existing catalogue entry for a data set from a 'search list' that is generated using any one of the following search criteria:

   7.6.1. The unique identifier of the data set.
   7.6.2. The name of the data set.
   7.6.3. A search string that is included in the description of the purpose of the data set.
   7.6.4. A search string that includes some or all of the keywords that represents the domain (or business area) covered by the data set.

[In the original document non-functional requirements (look and feel, usability, performance, operational, maintainability, portability, security and legal requirements) followed here. They have been omitted from this document.]