

## **Data Management Plan (DMP)**

### **Preamble**

The core infrastructural resource for this project will be the “Culture Box,” developed and housed at ETS and overseen by team leader, Mohamed Cheriet, Canada Research Chair in Cloud Computing (see the CFI description in the budget section for further explication). One student from Prof. Cheriet’s team will be dedicated to data management. The primary benefit of the Culture Box and its location at ETS in partnership with McGill will be its interoperability with Compute Canada’s High-Performance Computing cluster, CLUMEQ, which is overseen by McGill University and housed at ETS. A similar arrangement between a large storage cluster and high-performance processing system exists for the Artificial Intelligence Lab in Groningen (see <http://www.ai.rug.nl/~lambert/Monk-collections-english.html>). This will allow us to dedicate our computing resources to the storage of the large image data sets and use the HPC systems for the data-intense image processing and network analysis.

### **Data definition:**

The data involved in this project can be categorized in the following ways:

1. Document-image collections
2. Transformed and processed representations data for each collection
3. Discovered relational data for each collection
4. Methods and models to transform document images and to extract the relations
5. Implementations and realizations of methods of (4)
6. Network algorithm library
7. Research articles and publications
8. Project progress reports and project self-evaluations
9. Project website

### **1. Document image collections**

The primary data for this project consist of the four database collections, two of which are located at McGill University, one at Stanford, and one through an agreement with the HathiTrust. Those collections will be transferred to the central storage cluster in Montreal as well as the Monk storage cluster in Groningen for subsequent processing.

### **2. Transformed and processed representations data of each collection**

The data generated during the project by processing the four collections will be a key outcome of the project to be studied and analyzed by researchers beyond the project. These data will be hosted on the same storage cluster with open access to researchers.

### **3. Discovered relation data for each collection**

Similar to the transformed data, the relational data discovered will be a rich source of knowledge and information to be dug after the end of the project to discover high level knowledge about the collections, and even possible new approaches to their understanding. Because of the hierarchical nature of these data, it will be stored and access at various scales of complexity. Open access to these data will also be provided to researchers beyond this project.

#### **4. Methods and models to transform document images and to extract the relations**

The transformed data and relations will be the result of intensive research on related modeling, and will be associated to various analytical methods developed during the project. The knowledge related to these models and methods will be set open to the public, first, through disseminated research articles and, second, as unpublished research reports.

#### **5. Implementation and realization of methods**

The methods discussed in Section H of the proposal will be realized by implementing proper code. Although the team will leverage available open source research computing software, all code developed by the project team will be published as open source packages.

#### **6. Network algorithm library**

The size of the literary networks for this project requires that a specialized software library be used to infer the literary networks and compute the various features of interest to this initiative. The Zen Network Library (<http://zen.networkdynamics.org>), an open-source software library authored by team member Derek Ruths will be used for all network analytical tasks. Where new algorithms are developed, they will be made available through the Zen Network and the project website.

#### **7. Research articles and publications**

The main means of communication to the public and research community will consist of two primary formats: the publication of scientific articles and talks and the visual representation of the literary networks along with accompanying quantitative metrics. The network visualizations will be hosted on the project website as will short abstracts about the significance of the findings.

#### **8. Project progress reports and project self-evaluations**

We will undertake progress reports after each of the three stages of the project in each year of the grant. Self-evaluations in the form of lessons-learned reports will also be set open to the public both during and after the project. These will be important resources for future research to aid in the processual complexity of such interdisciplinary collaborations.

#### **9. Project website**

Synchromedia Lab will provide space for the project on its website that will host: project reports, blog posts, published articles and papers, visual representations of the networks, and links to data repositories. The web space will serve as a permanent host of the project beyond the grant time frame. As mentioned above, access to the project data will also be open to researchers via the storage cluster.