This is the first draft of a reference document defining the **technical infrastructure/data management** to be used for "The Congruence Engine: Digital Tools for New Collections-Based Industrial Histories " TaNC-AHRC-funded project.
As a first draft it necessarily contains much uncertainty, either because policies have to be determined, or because issues are currently under-specified. Input from other team members and stakeholders of the project about these topics would be particularly useful.

**** A work in progress. Please do not quote. Comments welcome. ****

The Congruence Engine: technical infrastructure and data management

v.0.1
Authors: Anna-Maria Sichani , Jamie Unwin, John Stack, Arran Rees

Overall Goal

Goal of this document is to describe the overall Congruence Engine's technical infrastructure strategy, including hosting, storage and preservation provision for the project's documentation, datasets and outputs.

Platforms/Services employed

GitHub

A public version of the CE GitHub Repository will be used for all public project-generated data/outputs, including

- datasets *
- prototypes' code
- project's public-facing documentation and reports (in line with the Publishing WG guidelines, TBC)
- training material

*Project partners' datasets are normally shared with the project under open licences such as CC0 - public domain data or CC-BY-SA (Attribution-Share Alike). Under these licences, data can be further processed and openly published in our CE Github repo. Where the data provider specifically claims a different copyright for their data, their dataset will remain in the **private Github repo.** Licensing should be made clear for all datasets.

A private GitHub Repo will be used mainly for datasets that are without open licences and for raw data from partners.

Amazon Web Services (AWS)

We do not plan to use this service in the first place , as this would require resources we do not intend to invest in, but we consider it as an option if and when the data demands it (while experimenting with ML).

SMG public-facing website

Project’s prototypes, data-related outputs and project-specific updates will be hosted in a dedicated-SMG website https://www.sciencemuseumgroup.org.uk/project/the-congruence-engine/ . This will be the central place for the project showcasing and communicating its findings in an accessible way, following also TaNC- AHRC requirements for publicly accessible outputs.

Zotero

A Zotero library has been set up, currently having a number of sub libraries, to store project’s bibliography.

Basecamp

Basecamp , currently hosted by MadLab (500GB storage included in our fee) is used for day-to-day communication of team members, general storage, updates and networking. A backup strategy of its contents must be in place.

OneDrive and GoogleDrive/docs

Cloud services such as OneDrive and GoogleDrive are used in everyday communication mainly for sharing documentation and facilitating meetings and discussions. Some of these documents will end up as project’s documentation hosted in the Public GitHub repo.

- Back up mechanisms for all the above must be in place.

system/platform	use	owner
GitHub	Public access ● datasets ● prototypes’ code ● project’s public-facing documentation and reports (in line with	JS, JU, AMS

	the Publishing WG guidelines, TBC) • training material Private - datasets with closed access	
Amazon Web Services (AWS)	Long-term storage Back up	JU
SMG repository	Project’s prototypes data-related outputs	JS, JU
Zotero	Project’s bibliography	DW, AMS
Basecamp	Day-to-day communication Events/meetings calendar Team networking	AR, AC, TB, NB,RT
OneDrive and GoogleDrive/	Day-to-day communication & documentation	All