RECODE: Towards a next-gen collections management system as part of the institutional and community data ecosystem

Steen Dupont[‡], Matt Woodburn[‡], Ben Collier[‡], Lucy Ellis[‡], Gerry Hey[‡], Tzy-Ting Hsu[‡], Mike Sadka[‡], Dave Smith[‡]

‡ Natural History Museum, London, United Kingdom

Abstract

Since 2020, the Natural History Museum, London (NHM) has been running the RECODE (Rethinking Collections Data Ecosystems) programme, an initiative that will provision a more open, manageable, configurable and interoperable collections management system (CMS) for the museum. With the overall aim of going live with an initial version of the new CMS by 2025, the first phase of defining a platform-agnostic set of high-level requirements and selecting a new technology partner and platform is nearing completion. The requirements, conceptual data models and other procurement documentation are shared openly through the Open Science Framework (OSF) platform so that any material may benefit and elicit feedback from the wider natural sciences community.

RECODE has strived to ensure that our new supplier and technology platform will be well positioned to deliver on the wider vision for community data interoperability, sharing and annotation. Through this presentation, we hope to continue our engagement with the global community by introducing our vision and describing our efforts to ensure that data sharing through technical interoperability and data standards are core features of the new solution.

As a digital representation of the collections and related processes, events and transactions, a CMS is an essential tool for many natural science collections, replacing

systems that were first analogue and paper-based, and later often distributed across multiple siloed, unstandardised, and unconnected files and databases. Consolidating that data and functionality into coherent, centralised application (as was first achieved at the NHM in 2002) facilitates more effective management of, and access to, both the physical collections and the data describing them. This consolidation also enabled the construction of a core collections data ecosystem within the museum, linking the CMS with frozen collections, providing some basic process for ingestion from digitisation workflows, and setting up a pipeline to offer up data to the NHM Data Portal for publication to the community (Fig. 1).
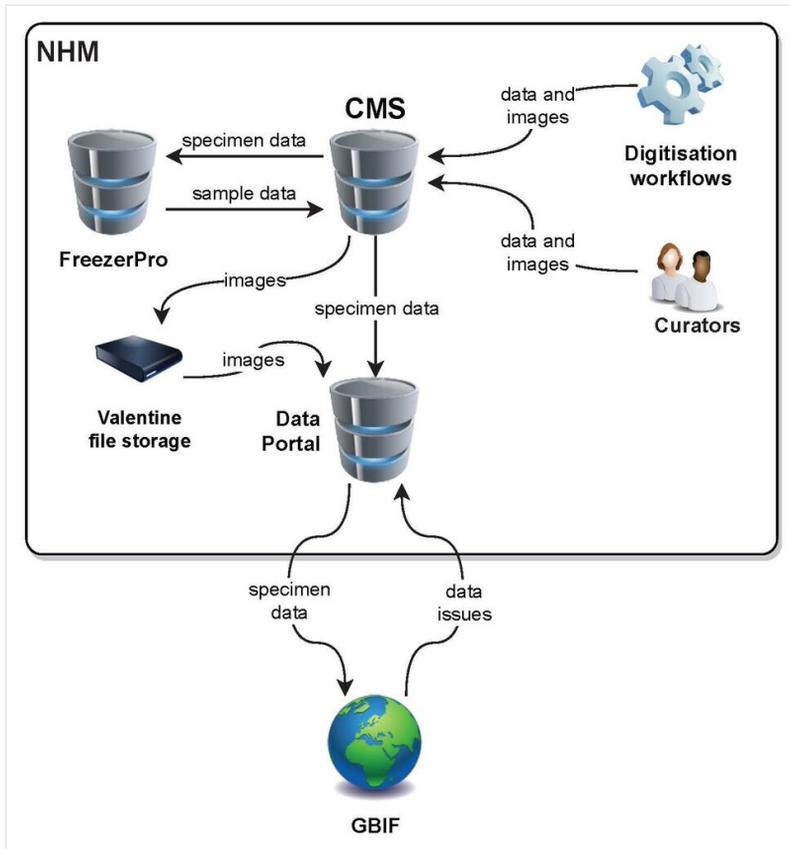

Figure 1.

A simplified representation of current interoperability within the NHM collections data ecosystem.

Although an important step on the path, the bespoke nature of these integrations, in part due to technical limitations in the CMS platform for importing and exporting data at scale, have limited further progress in this area. Even just within the museum's suite of science and collections data platforms there is a range of further potential integrations around the

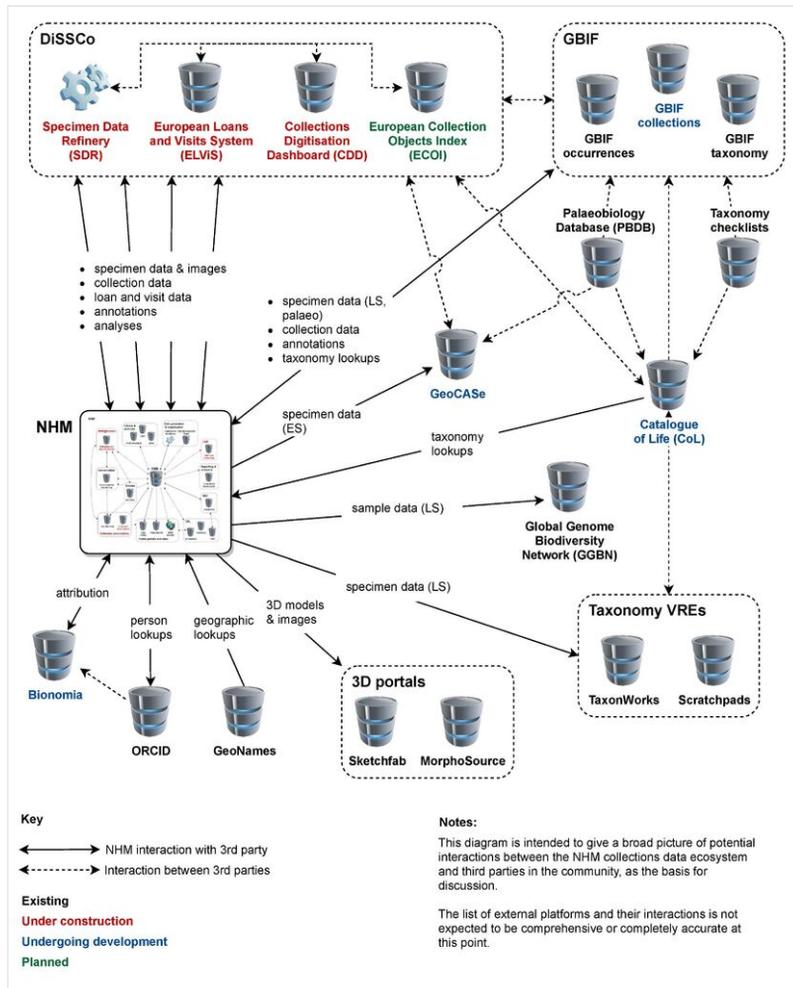CMS that could add considerable value in streamlining processes and joined-up decision support (Fig. 2).

Figure 2.

Potential interoperability for an enhanced NHM collections data ecosystem, supported by improved technical interoperability.

Modern technical capabilities, such as APIs, workflow capabilities and data models, dashboards and analytics, and integrated artificial intelligence (AI) and machine learning (ML) services, provide great potential for better management, sharing and exploitation of the data and the collections themselves. These capabilities, in particular those that support data interoperability, then open up much greater potential for positioning the institutional CMS within the wider external collections, biodiversity and geodiversity data ecosystem (Fig. 3). Not only does this offer much greater potential for using community-curated authorities, tools and services (e.g., Catalogue of Life, GeoNames, Bionomia and Wikidata

), but also closer integration with data aggregators and service providers such as the Global Biodiversity Information Facility (GBIF), Distributed System of Scientific Collections (DiSSCo), GeoCASe and Global Genome Biodiversity Network (GGBN), and opens up avenues for joining future initiatives like community data annotation.
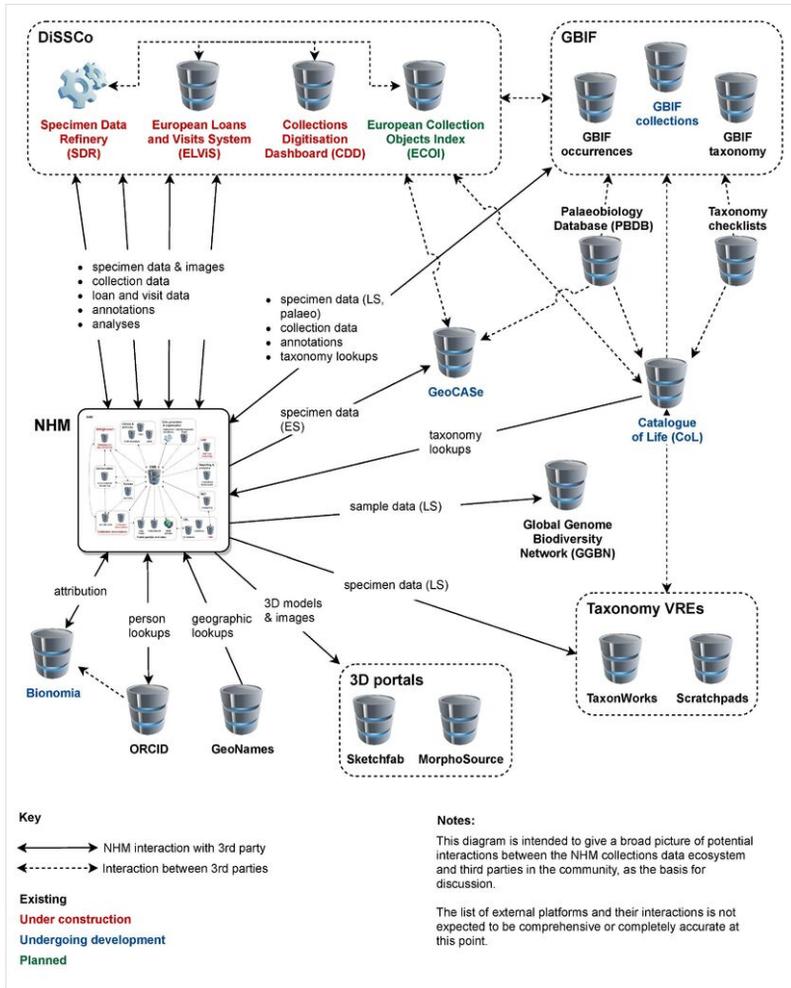

Figure 3.

Potential interoperability between the institutional data ecosystem (Fig. 2), with the CMS at the core, and community data platforms and services.

Over the past decade, the NHM has become increasingly aware that one of the major barriers to moving forward with our ambitions in this regard is outdated infrastructure and technology in the CMS marketplace, which has struggled to keep pace with the wider technology landscape. This realisation has driven the museum to consider more enterprise (and better resourced) technology sectors like Content Services Platforms (CSP). These platforms provide mature products that include these more cutting edge technical

capabilities, and tend to be highly configurable in order to be applicable across a wide range of domains. The onus, however, would be on us to design the data models and processes that would need to be configured within these platforms, which forms a major component of the RECODE programme. In this regard, both existing and emerging community standards and models like Spectrum, Darwin Core, Access to Biological Collections Data + Extension for Geosciences (ABCD+EFG), Latimer Core and the International Committee for Documentation Conceptual Reference Model (CIDOC CRM) are vital and will be used heavily to inform this work. Throughout the RECODE process, NHM intends to remain focused on the bigger community vision, and by creating a more open, flexible and community-ready CMS with a stronger focus on interoperability, standards, data quality and data sharing from the outset, pioneer a potential new CMS approach that may benefit others as well as ourselves.

Keywords

CMS, interoperability, data standards

Presenting author

Steen Dupont

Presented at

TDWG 2022