

## Global Identification for Unstructured Disparate Data Artifacts

### **Disclosure Number**

201102719

### **Technology Summary**

Many current data structures/databases consist of large collections of electronic documents, or articles, from different sources, both commercial and in-house. An individual document, for example a journal article, may consist of a number of items and be in different formats such as pdf, doc, docx, XML, text, proprietary, etc. It is expected that from multiple data sources, the same document may appear more than once. This invention disclosure addresses this issue by applying a Global Identification strategy and using disambiguation techniques to identify and group identical documents.

### **Inventor**

ABERCROMBIE, ROBERT K  
Computational Sciences & Engineering Div

### **Licensing Contact**

SIMS, DAVID L  
UT-Battelle, LLC  
Oak Ridge National Laboratory  
Rm 124C, Bldg 4500N, MS: 6196  
1 Bethel Valley Road  
Oak Ridge, TN 37831

Office Phone: (865) 241-3808  
E-mail: [SIMSDL@ORNL.GOV](mailto:SIMSDL@ORNL.GOV)

Note: The technology described above is an early stage opportunity. Licensing rights to this intellectual property may be limited or unavailable. Patent applications directed towards this invention may not have been filed with any patent office.