Data wrangling for effective deduplication
Organizations, contractors and third parties generate vast amounts of files and data during the lifecycle of any given project or asset. Information generated includes data spanning conceptual design, seismic, geological, engineering, detail design, construction, as built, operational, regulatory and decommissioning information.

Many global organizations are taking a fit-for-purpose approach to maintaining information and data. In particular, identifying true source information may seem daunting when faced with hundreds of thousands of files maintained in disparate systems and locations.

**Organizational challenge**

Over the years, this information has been maintained within structured and unstructured environments and shared with key stakeholders and company interfaces. Additionally, some information may be acquired and migrated from other systems and companies. As a direct result, storage costs and man-hours that go into searching and retrieving true source data are impacted significantly. Some of the challenges organizations face today include:

- Multiple duplicates stored over disparate locations.
- End-users accessing and utilizing incorrect revisions and not the true source, thus contributing to overall safety and risk.
- True source files are assigned with incorrect system attributes, and therefore become false duplicates.
- Users utilizing personal drives to maintain and share information without an audit trail.

**Data wrangling services**

Data wrangling services are ideally situated to assist with the most complex activities relating to deduplication and identification of true source files and data. A mandatory element of deduplication is the ability to find “near” duplicates rather than just exact copies. Our data wrangling features have been developed based on our intimate understanding of the inherent complexities associated with the domain.

Organizations require a deduplication solution that is tailored to their needs. It should encompass multi-levels of deduplication to eliminate redundant information and create a single source of the truth. Deduplication activities allow for the identification, verification and isolation of duplicate files, which can significantly reduce the amount of ROT (redundant and obsolete) information stored within areas, such as:

- Information management systems
- Archive areas: Legacy files and data which have lain dormant for years
- Acquired information stored within hard drives
- Unstructured personal/collaboration areas

Figure 2: Data wrangling services
Where files are not assigned with document numbering, revision coding or system attribution, organizations may require further interrogation, such as:

- Exact match
- Near match techniques such as:
  - Utilizing fuzzy matching images to find only near-match duplicates
  - Utilizing key metadata schemas (e.g., revision, creation date) to ensure near duplication is not simply a revision

Organizations need to ensure that positive duplicates are quarantined with the source master returning to the organization's target system. It is also important to compare any new file to all existing non-duplicate files in an optimized manner. This helps to ensure correct duplicate identification throughout.

**The result**

The end goal for any organization is to ensure that true source files and data are identified, maintained and easily accessible, while restricting unwarranted access and retaining the integrity of all information and related data.

One true source of information needs to be maintained and easily accessible to maximize the organizational activities surrounding any given project, asset, or exploration, while reducing the cost of maintaining redundant and duplicate information.

**Data wrangling approach**

Data wrangling services can manage the deduplication of files and metadata by utilizing several deduplication techniques:

- Identification of files and data using image comparison and text vector techniques
- Business rules for automatically determining which near duplicate image should be the master

**About the author**

Janine Murray  
Principal Consultant, Wipro Limited.

Janine Murray is an IM Consultant with over 15 years of experience in the O&G industry. She has extensive FE/Operations and Major Capital Project (MCP) Information Management experience. Janine also brings extensive experience with IM brownfield modifications, greenfield enhancements, MCP joint ventures, closeout, and MCP handover to Operations. Additionally, she is experienced in document cleaning and data extraction techniques for digitizing O&G legacy assets.

She can be reached at: janine.murray@wipro.com
Wipro Limited
Doddakannelli, Sarjapur Road,
Bangalore-560 035, India
Tel: +91 (80) 2844 0011
Fax: +91 (80) 2844 0256
wipro.com

Wipro Limited (NYSE: WIT,
BSE: 507685, NSE: WIPRO) is
a leading global information
technology, consulting and
business process services
company. We harness the
power of cognitive computing,
hyper-automation, robotics,
cloud, analytics and emerging
technologies to help our
clients adapt to the digital
world and make them
successful. A company
recognized globally for its
comprehensive portfolio of
services, strong commitment
to sustainability and good
corporate citizenship, we
have over 175,000 dedicated
employees serving clients
across six continents.
Together, we discover ideas
and connect the dots to build
a better and a bold
new future.

For more information,
please write to us at
info@wipro.com