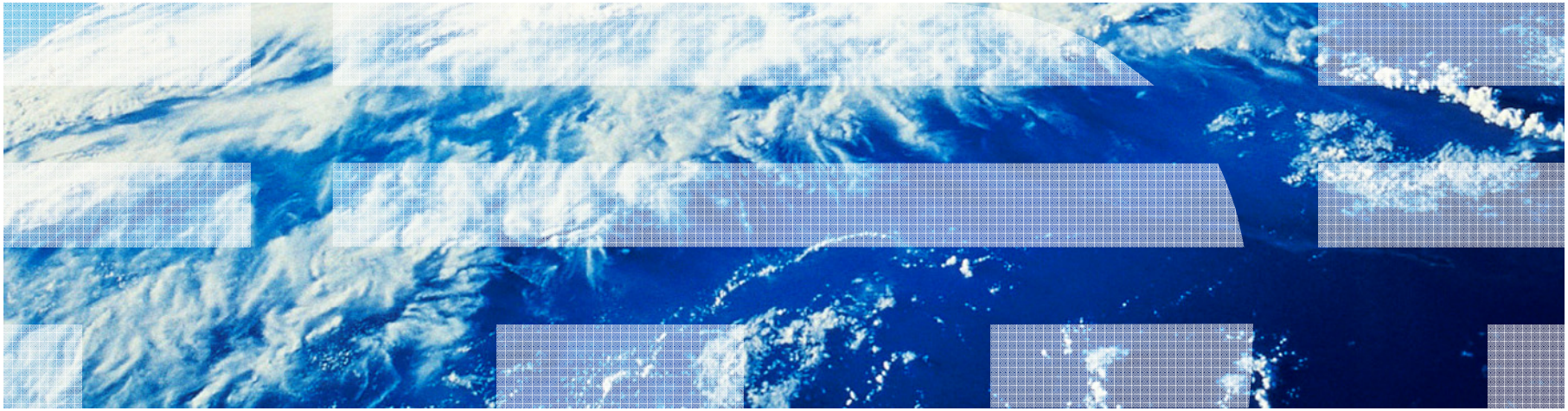
Karin Murthy, Deepak P, **Prasad M. Deshpande**, Sreekanth L. Kakaraparthy,
Vedula T. Surya Sandeep, Vijaya K. Shyamsundar, Sanjay K. Singh

IBM

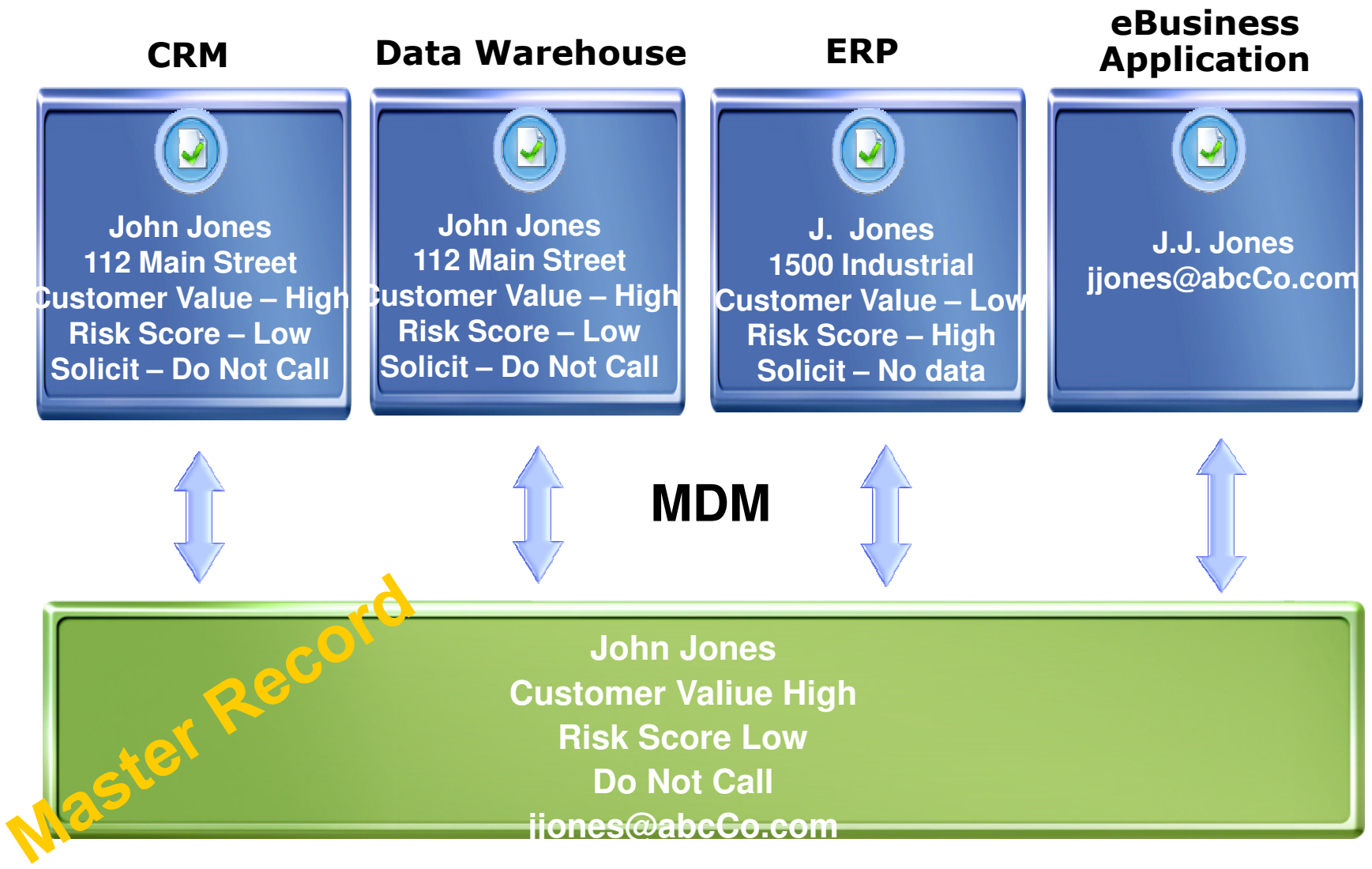
Content-Aware Master Data Management



MDM

- Master data management (MDM) indispensable for any enterprise to receive a
  - trusted,
  - integrated view
  - of all party-related information

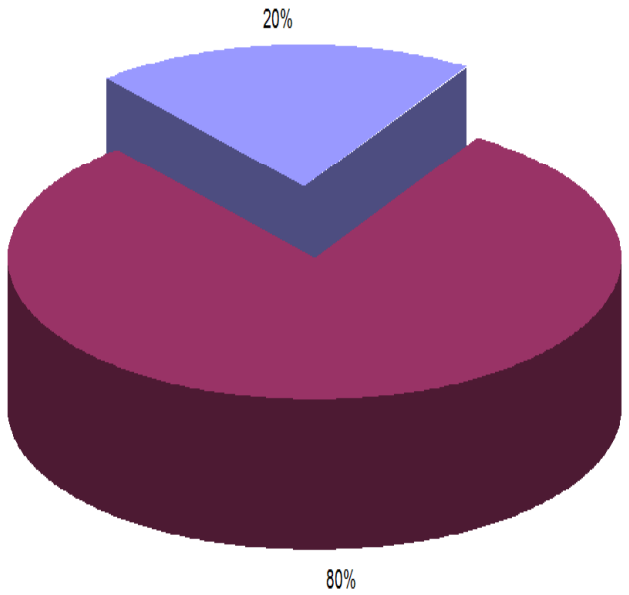
- For example, MDM provides a means to link data from various structured data sources and generate one integrated master record for each customer



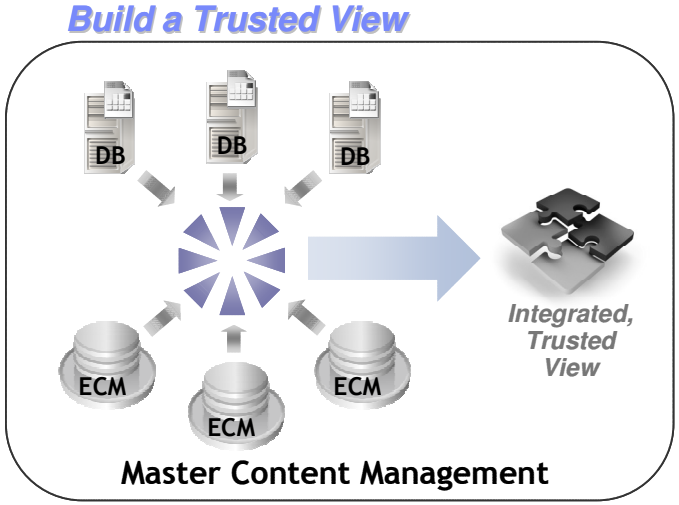
Business Problem – Integrating Unstructured Data Sources

- However, an estimated 80% of enterprise information is **unstructured**

- For example, large amount of valuable party information stored in the form of documents inside Enterprise Content Management (ECM) systems



Business Problem (continued)

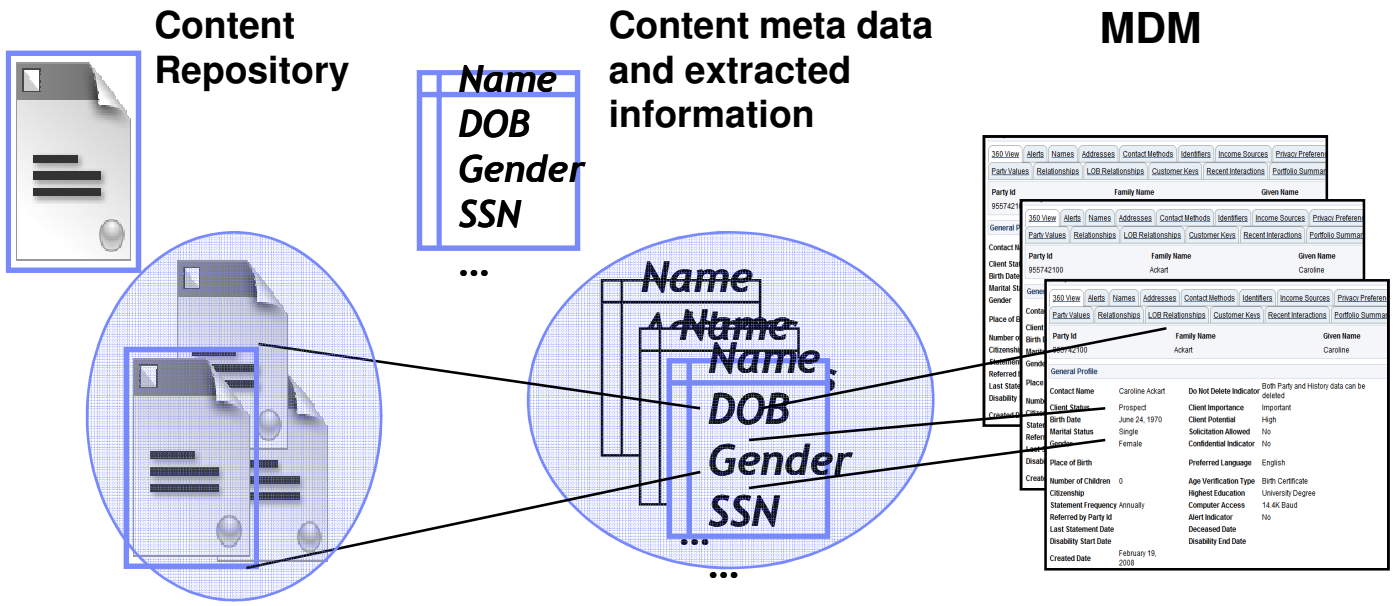


- InfoSphere Master Content Server (MCS)
  - bridges the gap between MDM and ECM
  - allows enterprises to link documents with existing master data records

- MCS has the following gaps
  - Unaware of document content
    - *documents are associated with the same entity based on metadata attributes alone*
    - *information contained in document is not added to master data record*
  - *No support for a "master" content*
    - *multiple versions or copies of content may exist*
  - *No validation of content*
    - *No relation between meta-data and actual content*

Making MDM Content-Aware

- Use content analytics to extract valuable information from each document and enrich its metadata
- Enhanced metadata enables
  - MCS to more accurately link content to master data
  - each master data record to be more comprehensive



Sample Application

- Staffing and Hiring
- Documents
  - CV, Cover letter, Reference Letters, Transcripts
- Useful information in the documents
  - name, phone, number, address, birth data, education, and employment history

- Uses of Content Aware MDM
  - Automatically populate the document metadata
  - Identify duplicate entries
  - Link with the master data to enable filtering of candidates

Use Case 1: Recognize errors in meta data

Local Entity	Document		Meta Data		Extracted Data			
ID	ID	Type	First Name	Last Name	First Name	Last Name	Student ID	Email
E1	doc1	CV	Ben	Doe	Ben	Doe		b.Doe@gmail.com
E1	doc2	Application	Ben	Doe	Ben	Doe	12345	b.Doe@gmail.com
E1	doc3	Application	Ben	Doe	Tom	Smith	9999	tom@yahoo.com

Doc3 is wrongly associated with party E1, but actually belongs to party E3. Suggest update of meta data in FileNet?

Use Case 2: Detect master content

Local Entity	Document		Meta Data		Extracted Data			
ID	**ID**	**Type**	**First Name**	**Last Name**	**First Name**	**Last Name**	**Student ID**	**Email**
E3	doc5	CV	Tom	Smith	Tom	Smith		
E3	doc6	CV	Tom	Smith	Tom	Smith		tom@yahoo.com

CV in doc6 is probably more relevant than CV in doc5.

Use Case 3: Detect suspect duplicate parties

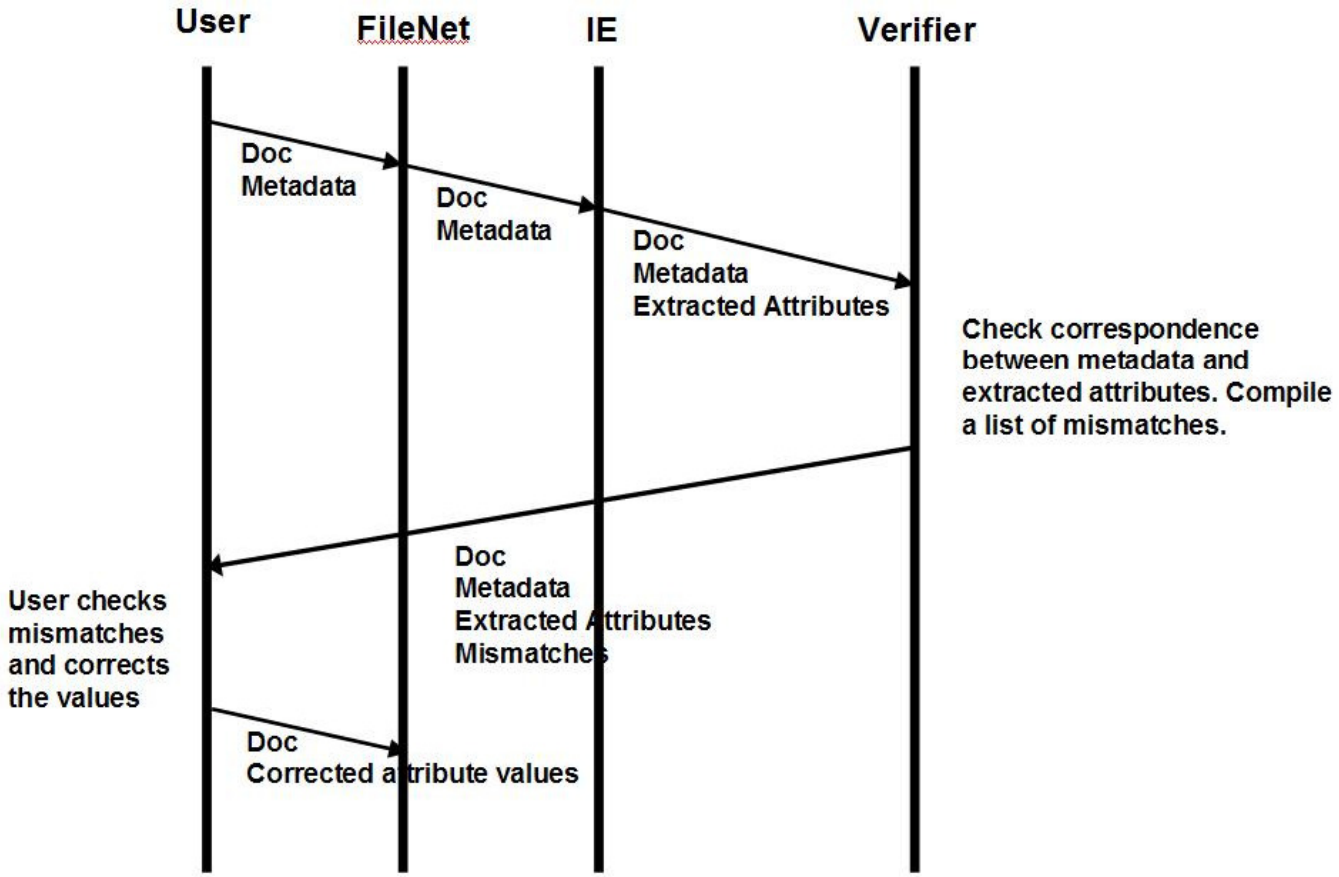
Local Entity	Document		Meta Data		Extracted Data			
ID	ID	Type	First Name	Last Name	First Name	Last Name	Student ID	Email
E1	doc1	CV	Ben	Doe	Ben	Doe	12345	b.Doe@gmail.com
E1	doc2	Application	Ben	Doe	Ben	Doe	12345	b.Doe@gmail.com
E2	doc4	CV	Benjamin	Doe	Benjamin	Doe	12345	b.Doe@gmail.com

Party E2 is with high likelihood a duplicate of party E1. Merge E1 and E2?

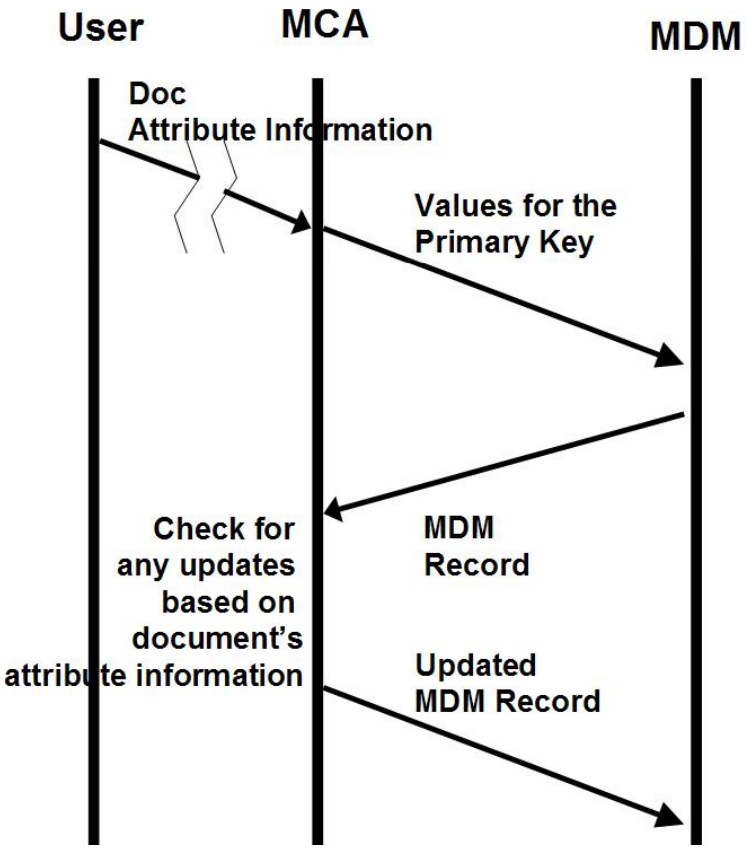
Components

- MDM, ECM
- Metadata Validator
  - Validating whether extracted information matches available metadata.
- Master Content Updater
  - Updating MDM with additional information available due to the upload of a document in ECM.
- Information Extractor
  - Responsible for extracting relevant information from unstructured documents.
  - Based on System T and AQL

Metadata Validator



Master Content Updater



High-precision Information Extraction

- Need high-precision annotators to deliver trusted data to MDM
- Rule-based annotators shown to achieve high accuracies
- Propose two solutions to further enhance accuracy

Utilize Available Metadata

Alg. 1 ***Enhanced Information Extraction***

Input 1: $\{< a_i, v_i >\}$, available attribute-value pairs
Input 2: $\{e_j\}$, attributes for extraction

1. $\forall a_i$
2. run the annotator for a_i, and extract
3. all possible values as V_i
4. $given = \bigcup_i v_i$
5. $all = (\bigcup_i V_i) \cup given$
6. $\forall e_j$
7. run the annotator for e_j, and extract
8. all possible values as S
9. $\forall c \in S$
10. $closest_s = argmin_{v \in all} text\_distance(s, v)$
11. $if(closest_s \notin given)$
12. $S = S - \{s\}$
13. $s' = argmin_{s \in S} text\_distance(s, closest_s)$
14. set s' as chosen value for e_j

Dear Biju,
*This is with respect to my recent application (reference number **9456734231**). Sorry to hear that you had trouble contacting my old employer. You should be able to reach the correct representative in the HR department of XYZ at **9876543211**. His name is Babu.*

Regards,
Arun
Software Engineer,
XYZ Inc., Bangalore – 74
9876456789

Occurrence	Distance from Arun
9456734231	34
9876543211	5
9876456789	5

Incorporate Selective User Feedback

- Associate confidence scores with both final annotations as well as intermediate results
- Use provenance framework provided by rule-based IE systems to update confidence scores appropriately

Experimental Evaluation

- Results for Indian resume data

Annotator	Precision	Recall
Person Name (generic)	33	32
Person Name (with metadata)	92	48
Phone Number (generic)	100	80
Phone Number (domain-specific)	100	92
Email (generic)	100	100
Date of Birth	100	92
Highest Qualification	96	96
Year of Qualification	100	96
Current Employer (generic Org annotator)	91	76
Current Employer (domain-specific Org annotator)	100	88
Years of Experience	95	80

Conclusion

- Can harness content for master data management
  - Possible to extract reliable structured information from content
- Used to link with other master data for an entity, to detect master content, to enhance detection of duplicate entities, and to validate metadata associated with documents.
- Content Aware MDM is possible