Data Integration, Simplified

Luc Pezet, v.1

March 31, 2014

Definition 1. Let D be the set of all documents. Documents may belong to Source data or Target data.

Definition 2. Let $T \subseteq D$ be a set of integration results, i.e. the Target set. T is generally considered to be the final result of compiling multiple Source sets through successive integration steps.

Definition 3. Let $S \subseteq D$ be a set of Source documents considered for integration into a Target set.

Definition 4. Let $rep_i : S, T_i \mapsto \{0, 1\}$ such that for $x \in S, y \in T_i$,

$$rep_i(x, y) = \begin{cases} 1 & \text{if } x \text{ and } y \text{ represent the same document} \\ 0 & \text{otherwise.} \end{cases}$$

Definition 5. Let Υ_i be a function that creates a document in T_{i+1} from a document of S. Formally, let $\Upsilon_i : S \mapsto T_{i+1}$ such that for $x \in S$,

$$\Upsilon_i(x) = y \mid y \in T_{i+1} \wedge rep_{i+1}(x, y) = 1$$

Definition 6. Let $\gemini_i$ be a function that modfies a document from T_i using a document from S into T_{i+1}. Formally, let $\gemini_i : S, T_i \mapsto T_{i+1}$ such that for $x \in S, y \in T_i, rep_i(x, y) = 1$

$$\gemini_i(x, y) = z \mid z \in T_{i+1} \wedge rep_{i+1}(x, z) = 1$$

Definition 7. Let $\capricornus_i : S \mapsto T_{i+1}$ the integration function of $x \in S$ with documents in T_i into T_{i+1}, such that $\forall x \in S$,

$$\capricornus_i(x) = \begin{cases} \Upsilon_i(x) = z & \text{if } \forall y \in T_i,\ rep_i(x, y) = 0. \\ \gemini_i(x, y) = y' & \text{if } \exists y \in T_i,\ rep_i(x, y) = 1. \end{cases}$$

Lemma 1. *The function $\capricornus_i$ is idempotent if and only if the subsequent integrations leads only to the $\gemini$ functions. Formally,*

$$\capricornus_i(x) \textit{ idemptotent } \iff \capricornus_{i+1}(x) = \gemini_{i+1}(x, y)$$

Proof. If for $x \in S$ and $\forall y \in T_i, rep_i(x, y) = 0$,

$$\Longrightarrow \capricornus_i(x) = \Upsilon(x) = z \tag{1}$$

$$\Longrightarrow \exists z \in T_{i+1} \mid rep_{i+1}(x, z) = 1 \tag{2}$$

$$\Longrightarrow \capricornus_{i+1}(x) = \gemini_{i+1}(x, z)$$

□

If for $x \in S$, $\exists y \in T_i \mid rep(x, y) = 1$,

$$\implies \capricornus_i(x) = \gemini_i(x, y) = y' \tag{3}$$

$$\implies \exists y' \in T_{i+1} \mid rep_{i+1}(x, y') = 1 \qquad \implies \capricornus_{i+1}(x) = \gemini_{i+1}(x, y) \tag{4}$$