careevolution
HEALTHCARE TECHNOLOGY

Identity Management

An overview of the CareEvolution RHIO Technology Platform's Identity Management (record linking) service

Identity Management

Identity management, or patient record linkage, is an important and necessary step in enabling the exchange of clinical information between the healthcare information systems of disparate hospitals and clinics. The CareEvolution RHIO Technology Platform incorporates a Record Locator Service (RLS) built on top of industry standard record linking techniques providing a set of advanced features that improve security and link specificity.

At first blush the RLS requirements look very similar to what an off the shelf MPI solution can offer. However there are significant security and privacy issues involved with a regional deployment that must be addressed. The CareEvolution Record Locator Service implements robust record linking techniques inspired by industry and academic research and designed to handle unique RHIO considerations.

Securing Identity Management - The Privacy Mandate

RLS - The Traditional Weak Link in the RHIO Security Chain

CareEvolution believes that the centralized store of accessible demographic information employed by most RHIO implementations creates an unacceptable security risk for any RHIO. Data aggregation accentuates three critical risk factors that increase the potential that sensitive information will be improperly disclosed:

- First, data aggregation increases the value of the centralized store creating a lucrative target for potential attackers.
- Second, it increases the number of entities that legitimately should have access to the central store; this in turn increases the number of avenues that can be compromised by attackers.
- Third, a centralized store of sensitive data can become a valuable resource that may be susceptible to political pressure for legalized access by interests claiming a need to know. A concerted effort by the government to obtain data from the large Internet search engines is a compelling example of this third risk factor.

Blinded Record Linkage - The Solution

Methods must be deployed that can strongly secure this centralized data store. The CareEvolution RHIO Technology Platform provides a solution for this challenge. The platform achieves a secure, performant solution to record linkage in the distributed system by using a **blinded directory** for centralized demographic data. A set of techniques are implemented to cryptographically (one-way) hash the aggregated data to ensure that patient demographic data stored in the centralized index is unrecoverable. There are two direct results of hashing the centralized index :

- **World Class Security -** From any plaintext string (i.e. "Smith"), a one-way hashing algorithm can quickly produce a long sequence of numbers (a "hash"), which represents the string "Smith". However, to take this hash and reverse the algorithm to arrive at "Smith" would require years of computation, hence the term "one-way" hash.

- **Record Linking Challenge** – Since hashes of similar strings, such as "Smith" and "Smit" yield drastically different number sequences, the very process of hashing renders traditional approximate record linking techniques inoperable. As a result, all contemporary providers of MPI or identity management solutions have avoided the formidable technical challenges posed by a crypto-hashed central directory. While this may have been acceptable when such solutions were intended to be implemented behind the security firewalls within an institution, **we believe that extending a non-hashed centralized repository of demographic information across a region, let alone the country, poses an unprecedented and unwarranted privacy risk.**

Beyond Security

In addition to the critical security component there are several other important requirements that an RHIO deployment demands.

- **No Mistakes -** There must be a near-zero false positive link rate. The worst case scenario for an RHIO record linking system is to incorrectly link patients and move or display data on the wrong patient record. The CareEvolution record linking algorithms are tuned to near 100% specificity to prevent false positives.

- **Automated -** The platform should leverage record linking techniques that can perform the vast majority of the linking activity. Where manual review is appropriate CareEvolution provides a streamlined user interface so that administrators can quickly review and accept links.

- **Real-time -** The system should perform in near real-time so that clinically relevant information from remote institutions can be incorporated in existing workflows.

CareEvolution Record Linking Fundamentals

Some of the basic components in the design of the CareEvolution record linking system include data standardization and configurable deterministic and probabilistic linking strategies.

- **Standardization -** Demographic information is "cleansed" so that comparing this information will yield meaningful results. Casing, white

space, special characters, nicknames, and configurable garbage values must be handled uniformly for each institution's record store.
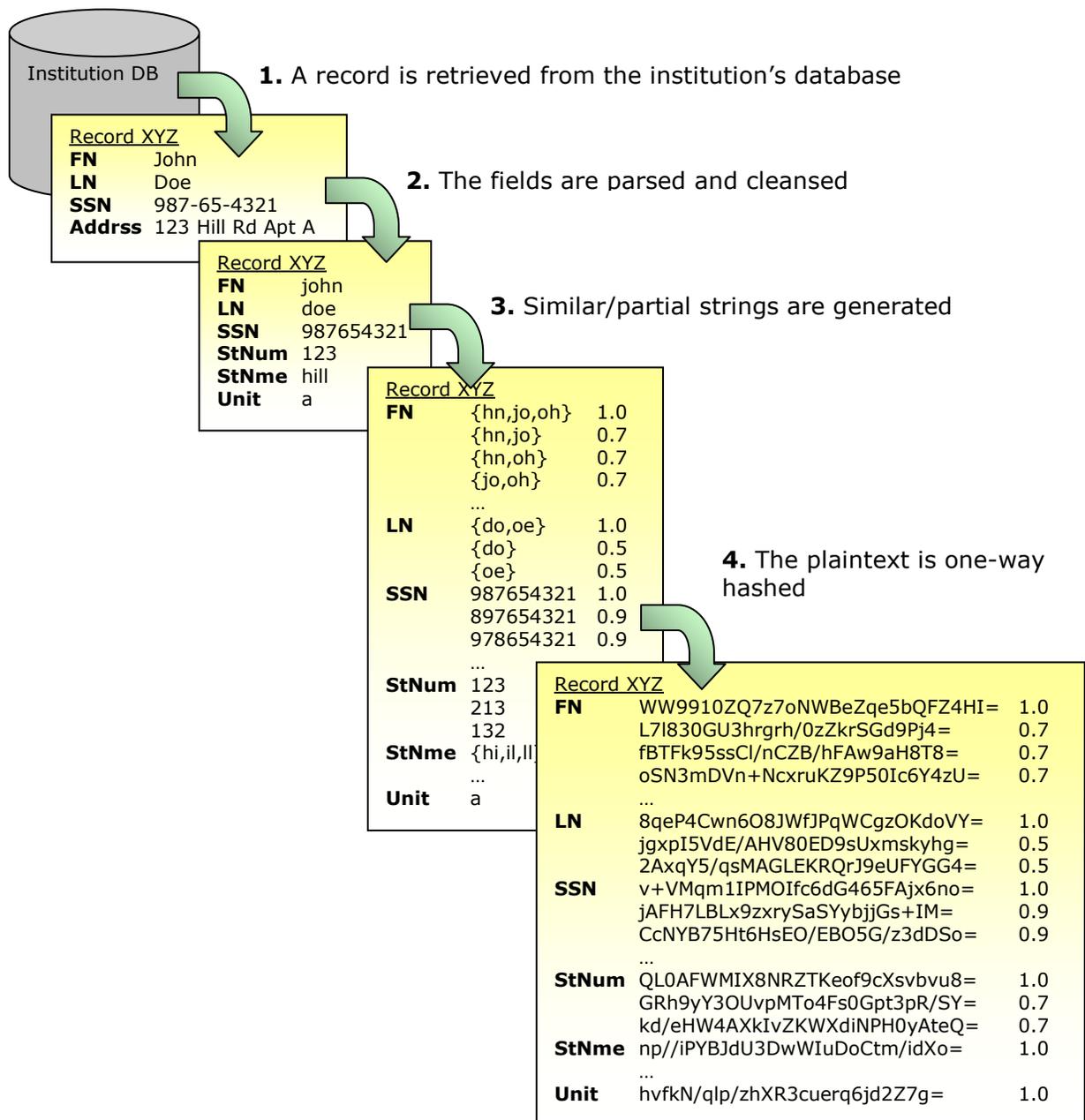
- **Record Linkage -** After the record has been standardized and transformed, record pairs are compared to determine their similarity. There are a range of established techniques that assist in this effort. Two primary categories include deterministic linking and probabilistic linking. Deterministic record linking employs a series of simple rules (SSN, DOB, and last name match exactly, for example) to determine linkage. While deterministic rules can be designed for high specificity, they still leave room for improvement. For example, in the rule described above, suppose the first name and gender varied greatly. These differences cast doubt on the validity of the link; it could be that one record has a mistyped SSN, or perhaps a husband and wife were both entered in the system with the same SSN for insurance reasons. Probabilistic record linking, on the other hand, associates statistical importance for agreement and disagreement to each identifier. Two records link if the sums of all the weighed agreements outweigh the disagreements. The CareEvolution record linking system allows for multiple configurable linking strategies that all help determine a record pair's final link status.

## CareEvolution Advanced Record Linking Techniques

CareEvolution's RHIO Technology Platform record linking system builds on the solid foundation of mainstream record linking systems with advanced features including identifier blindfolding, consistency checking, and an advanced human review feedback loop.

- **Blindfolded RL -** In traditional RLS models plaintext demographic information is centrally located to facilitate record linking. To preserve privacy, CareEvolution implements a blindfolded record linking system that cryptographically hashes record identifiers; obfuscating the information in such a way that comparisons can still be made but the original clear-text is irrecoverable. This provides the best of both worlds in that data can be freely shared for the purpose of record linking, but that data same data can not be read due to the nature of this one-way hash.

- **Blindfolded Approximate Matching –** Because the hashes of similar identifiers bear no correlation with each other preprocessing of the unencrypted identifiers must be done to allow for approximate matches with identifier hashes. Standardized demographic information is transformed before blinding to allow for approximate string matching. Approximate matching in this scheme is accomplished using a technique called bigramming. [1] Bigramming breaks up the source string into many derived strings. Each derived string is given a

similarity score that indicates how similar it is to the source. Two strings that have been bigrammed can then be compared by determining if they share a derived string. If so, the two derived similarity scores can be used to compute an overall "dice score." Using a bigramming technique to generate derived strings and then hashing these strings enables approximate, blinded identifier matching.

Institution DB

1. A record is retrieved from the institution's database

Record XYZ
FN	John
LN	Doe
SSN	987-65-4321
Addrss	123 Hill Rd Apt A

2. The fields are parsed and cleansed

Record XYZ
FN	john
LN	doe
SSN	987654321
StNum	123
StNme	hill
Unit	a

3. Similar/partial strings are generated

Record XYZ
FN	{hn,jo,oh}	1.0
	{hn,jo}	0.7
	{hn,oh}	0.7
	{jo,oh}	0.7
	...	
LN	{do,oe}	1.0
	{do}	0.5
	{oe}	0.5
SSN	987654321	1.0
	897654321	0.9
	978654321	0.9
	...	
StNum	123	
	213	
	132	
StNme	{hi,il,ll}	
	...	
Unit	a	

4. The plaintext is one-way hashed

Record XYZ
FN	WW9910ZQ7z7oNWBeZqe5bQFZ4HI=	1.0
	L7l830GU3hrgrh/0zZkrSGd9Pj4=	0.7
	fBTFk95ssCl/nCZB/hFAw9aH8T8=	0.7
	oSN3mDVn+NcxruKZ9P50Ic6Y4zU=	0.7
	...	
LN	8qeP4Cwn6O8JWfJPqWCgzOKdoVY=	1.0
	jgxpI5VdE/AHV80ED9sUxmskyhg=	0.5
	2AxqY5/qsMAGLEKRQrJ9eUFYGG4=	0.5
SSN	v+VMqm1IPMOIfc6dG465FAjx6no=	1.0
	jAFH7LBLx9zxrySaSYybjjGs+IM=	0.9
	CcNYB75Ht6HsEO/EBO5G/z3dDSo=	0.9
	...	
StNum	QL0AFWMIX8NRZTKeof9cXsvbvu8=	1.0
	GRh9yY3OUvpMTo4Fs0Gpt3pR/SY=	0.7
	kd/eHW4AXkIvZKWXdiNPH0yAteQ=	0.7
StNme	np//iPYBJdU3DwWIuDoCtm/idXo=	1.0
	...	
Unit	hvfkN/qlp/zhXR3cuerq6jd2Z7g=	1.0

- **Human Review -** Automated record linking requires a tradeoff between sensitivity and specificity. Even with advanced record linking techniques, the ultra-high specificity required by the RHIO platform means that some actual links will be left in a possible state. So, CareEvolution has implemented an Identity Management tool that allows administrators to weigh in, upgrading possible links to definite, and thus allowing the flow of clinical data between those two records. This rich interaction between the Record Locator Service and system users enable the CareEvolution RHIO Technology Platform to achieve high specificity without sacrificing sensitivity.

- **Consistency Checking -** Because administrators are able to weigh in on the validity of links established by the Record Locator Service, the potential for set inconsistencies exists. Take, for example, three institutions. A user at Site A reviews a record and approves a link to Site B as well as a link to Site C. That is, there are links A-B and A-C. However, a user at Site B disapproves the link B-C. This inconsistent state is managed and escalated by the CareEvolution Record Linking System; disallowing the flow of clinical data until the inconsistency is resolved.
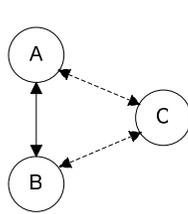
In a pair wise record linking process record C is found to be a possible match with A and C is found to be a possible match with B
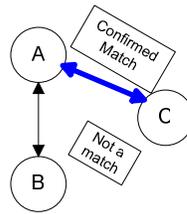
A human reviewer from record C's source institution wishes to confirm that there is a link from C to A but deny the link between B and C
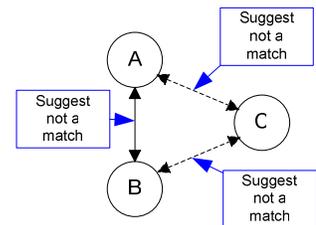
The consistency checker will not allow C to match with A and not match with B while A is a match with B. If A and B are the same person and A and C are the same person then B and C can't NOT be the same person. The reviewer at site C will be allowed to suggest the correct match state but a reviewer at site A or B must agree that A does not match B and effect the change to that link accordingly.

Summary

The need for very high specificity as well as appropriately high sensitivity challenges the health care community today. Additionally, the need for centralized demographic information in RHIOs presents a serious privacy risk. By leveraging state of the art record linking techniques, the CareEvolution RHIO Technology Platform is able to address these issues and provide secure linking with a near-zero false positive rate, while allowing human review to help find all possible links.

References

[1] Tim Churches and Peter Christen. Some methods for blindfolded record linkage. BMC Medical Informatics and Decision Making, 4(9), 2004
http://www.biomedcentral.com/1472-6947/4/9

About CareEvolution, Inc.

CareEvolution is a leading provider of secure interoperability solutions. Our RHIO platform offering is a robust Service Oriented Architecture (SOA) to enable RHIOs' heterogeneous underlying EMRs to "share" clinical information in a secure, reliable, and incremental manner. Distinct components such as Identity Management, Record Location, Clinical Data Integration, Audit & Log, Data Persistence, Visualization, Terminology, and Data Mining may be adopted piecemeal or as a comprehensive technology platform.

For More information please visit us at www.careevolution.com