FAIR Data Management: Workshop di Research Data Alliance (RDA) a Firenze

Silvia Vellani, Valentina Lepore

Nei giorni 14 e 15 novembre 2016, presso la Biblioteca Umanistica dell'Università di Firenze, si è tenuto un incontro di iniziativa e vocazione internazionale: il workshop *Fair Data Management*: *Best practices and open issues*, organizzato da RDA, the Research Data Alliance, in collaborazione con le Università di Torino, Milano, Bologna, Trento e Parma, e supportato dal CNR, l'Università di Firenze, AISA (Associazione Italiana per la Promozione della Scienza Aperta) e OpenAIRE (Open Access Infrastructure for Research in Europe). La sinergia fra le istituzioni interessate riflette l'ideale di condivisione e collaborazione alla base di RDA.

Ma cos'è RDA, qual è la sua visione? Quale la sua *mission*? Chi vi aderisce, e chi può aderire? La risposta a queste domande, in apertura, dà modo a Donatella Castelli di mettere a fuoco i temi di interesse delle due giornate[1].

RDA è una comunità di pratica e di ricerca internazionale che si identifica ed opera all'insegna dei cosiddetti principi FAIR: i principi guida definiti nel 2014 per una gestione dei dati che renda i contenuti informativi *Findable*, *Accessible*, *Interoperable* e *Re-usable*. A RDA può pertanto aderire, a vario titolo (lo dimostra il grafico dei membri individuati per professione di appartenenza: ricercatori, informatici, studenti, bibliotecari, docenti, amministratori delegati, giornalisti, editori, etc.), chiunque abbracci la causa dei dati di ricerca. E l'obiettivo principale del workshop è, in tal senso, suscitare un coinvolgimento più profondo e più coeso delle istituzioni, degli enti e delle università italiani. Castelli illustra le «recommendations and flagship outputs» di RDA, cioè i suoi prodotti e i suoi standards rappresentativi, ciascuno dei quali assegnato ad un *working group* di competenza specifica. Due esempi varranno a rilevare l'impatto complessivo dei prodotti RDA nella comunità scientifica: 1) «Basic Vocabulary of Foundational Terminology Query Tool», ossia un vocabolario condiviso, e volutamente non settoriale, di termini e definizioni apposite per gli oggetti digitali: per facilitare lo scambio e la comunicazione dei dati fra esperti di ambiti di ricerca differenti; e 2) «Scalable dynamic-data Citation Methodology», cioè un meccanismo scalabile che garantisca un'identificazione precisa e stabilizzata di dati soggetti a modifiche. I due esempi testimoniano senza dubbio molto bene l'obiettivo principe di RDA: l'abbattimento delle barriere che ostacolano la cultura del *data sharing*, siano esse disciplinari (1), o tecnologiche (2).

Si rivolge in modo preminente alla comunità delle biblioteche scientifiche anche LIBER (Ligue de Bibliothèques Européennes de Recherche): «la rete di biblioteche più estesa d'Europa», con più di 400 membri sul territorio. La direttrice esecutiva Susan Reilly ne ha descritto le pratiche e gli intenti per i dati di ricerca[2], in programmatico adeguamento ai principi FAIR e alle direttive del mandato europeo Horizon 2020 per l'Open Access alle pubblicazioni scientifiche e ai dati di ricerca2. Quello che LIBER persegue è un modello di conoscenza sostenibile, che assicuri la conservazione, l'accesso e la fruizione dei risultati della ricerca alla generazione presente e a quella futura: perché «il patrimonio e l'eredità culturale si costruiscono sulle informazioni digitali di oggi» – o, più precisamente: in ragione delle modalità oggi utilizzate (e dei costi ad esse relativi) per la loro raccolta, archiviazione e diffusione. I risultati della ricerca scientifica sono un bene pubblico secondo LIBER (e non solo, ovviamente), perciò i dati

[1] Presentazione dal titolo "Research Data Alliance Overview"

[2] Presentazione dal titolo "Collaborate to share"

devono essere condivisi e resi *open*.

Integra, tuttavia, questa prospettiva Andreas Rauber[3]. Ai fini della preservazione e della diffusione dei contenuti informativi, infatti, la condivisione è una condizione necessaria ma non sufficiente: occorre che i dati siano anche riproducibili. Ma in cosa consiste, esattamente, la riproducibilità di un contenuto informativo, e qual è il guadagno strategico che ne deriva? Un contenuto è riproducibile nella misura in cui esso si dà quale esito di un processo eseguibile anche da parte di terzi. Ciò significa che, allorché si pubblicano i risultati di una determinata ricerca, si debba contestualmente provvedere anche all'accessibilità ai dati in essa utilizzati e alla procedura di elaborazione degli stessi. O, in termini ancor più tecnici, occorre che il ricercatore dia riferimento, oltre che dei programmi e dei dati utilizzati, anche di codice sorgente e formato dati relativi, e che egli fornisca inoltre il collegamento ai *research data* archiviati in un *open repository*. I vantaggi che ne derivano sono molteplici e vari. Ma nella rassegna di Rauber i più degni di nota sono quelli della trasparenza e della possibilità di controllare i risultati (cioè di verificarli e di falsificarli), ciò che viene a interessare non solo la comunicazione dei suddetti risultati alla comunità scientifica, bensì il loro intrinseco contenuto di verità e i corrispettivi criteri di validazione: la posta in gioco, dunque, è di ordine anzitutto epistemologico.

Sul progetto Horizon 2020 entra nel merito Paola Gargiulo[4]. La Commissione europea ha sancito che dal 2017 il modello *open data* diverrà l'opzione di *default* in ambito di ricerca, e a questo proposito, con lo scopo di valutarne le strategie di attuazione ottimali, ha lanciato l'*Open Research Data Pilot*: un progetto pilota, appunto, che dal biennio 2014-2015 (e aperto ad ulteriori auto-candidature) ha selezionato e finanzia alcuni enti di differenti aree scientifiche affinché essi rendano accessibili – salvo giustificata riserva di *opt out* relativa a dati sensibili – i loro *research data*3.

Il punto di partenza, e il primo passo per candidarsi al progetto *Pilot*, è redigere un *Data Management Plan*, cioè un documento (non immutabile: esso potrà essere modificato e aggiornato in fasi successive) nel quale l'ente interessato deve: 1) descrivere il *set* di dati (nel loro contenuto e formato) che si vogliono creare, raccogliere e preservare; 2) dare riferimento degli standards che si useranno per la meta-datazione; 3) individuare il pubblico potenzialmente interessato ai dati raccolti; e 4) specificare le tecniche, le politiche e i costi di preservazione, accesso e riuso degli stessi. Tale redazione richiede, pertanto, competenze specifiche 'nuove', alla formazione e all'espletazione delle quali sono destinati, da una parte, appositi servizi online (linee guida certificate, webinars, e programmi di training) erogati, ad es., da OpenAIRE e Eudat, e, dall'altra, alcuni *free tools* in rete come il *repository* multidisciplinare Zenodo, e DMPonline elaborato dal Digital Curation Centre.

Sennonché resta inteso, in ogni caso – e nel corso del workshop Erik Shultes lo sottolinea[5] –, che le nuove competenze di cui sopra, e in generale quelle complessivamente necessarie alla produzione e alla diffusione di FAIR data, si poggiano interamente su tecnologie preesistenti.

Al vivace contesto europeo fa da contrappunto il ritardo italiano in quanto a *policies* istituzionali e direttive ministeriali in ambito di *Research Data Management*. Sulla questione fa il punto Paola Galimberti[6], la quale non si sofferma, però, solo sui punti deboli che tale ritardo comporta, ma anche sui punti forza. Se è vero, infatti, che la mancanza di un piano

[3] Presentazione dal titolo "Reproducibility challenges in computational settings: what are they, why should we address them, and how?"

[4] Presentazione dal titolo "OpenAIRE and Eudat services and tools to support FAIR DMP implementation"

[5] Presentazione dal titolo "FAIRness through a novel combination of Web technologies"

[6] Presentazione dal titolo " The Italian Universities RDM WG: tools and best practices"

a livello governativo può ridurre le opportunità di finanziamenti pubblici e togliere i vantaggi in termini di coesione e organicità dei risultati garantiti da un coordinamento centrale, è vero anche, per converso, che il vuoto rimasto lascia il campo libero alla formazione di *working groups* spontanei, auto-regolamentati e informali. E la collaborazione senza vertice ha i suoi vantaggi da arrecare alla causa RDM: rende più agevole e immediata la collaborazione orizzontale interna al gruppo di lavoro, e conta unicamente su collaboratori non obbligati da terzi, dunque altamente motivati.

Dopo la riflessione teorica, Lucia Mona passa ad un primo esempio di applicazione pratica di gestione dati FAIR[7]: ACTRIS-2 IA è un progetto del programma Horizon 2020 (quadriennio 2015-2019) che ingloba, integra e si innesta sulla collaborazione (*ante* 2011) dei progetti europei EARLINET, EUSAAR, CREATE e Cloudnet. L'obiettivo è osservare e interpretare i mutamenti e le tendenze della composizione atmosferica onde valutarne l'impatto nella stratosfera e nella troposfera, e dunque spiegare e controllare i mutamenti climatici. I campi di applicazione della tecnologia informatica ai *research data*, ad ogni modo, sono (e devono essere) tanti quanti gli ambiti di ricerca esistenti.

Ma Franco Niccolucci mette in rilievo le difficoltà maggiori che redigere un DMP per un progetto di ambito umanistico comporta[8]. Ciò è dovuto in primo luogo alla natura complessa e eterogenea dei *primary data* a soggetto, giacché, ad es., un manufatto archeologico richiede competenze e tecniche di digitalizzazione e di preservazione diverse da quelle necessarie per un manoscritto o una stampa.

Antonio Rosato presenta un altro progetto, di bioinformatica, incluso nel programma Horizon 2020[9]. Al riguardo Rosato ribadisce che i vantaggi di una condivisione dei *research data* si realizzano se e solo se ciascun ricercatore si preoccupa non solo dell'elaborazione del dato ma anche della sua *digital curation* lungo l'intero ciclo di vita del dato stesso: nel caso specifico, se e solo se ciascun ricercatore partecipe di INSTRUCT provvede alla preservazione dei dati che egli stesso immette nell'apposito *open repository* in *cloud storage* wwPDB (worldwide Protein Data Bank).

Sorge dunque la questione: in cosa consiste la *digital curation*? O, altrimenti: «*Data curator: who is s/he?*». Sulla base di un'indagine IFLA 2015-2016, Anna Maria Tammaro mette a fuoco le esigenze formative, le competenze e la peculiarità del *data curator*. Di primaria importanza, infatti, è comprendere se la *digital curation* si costituisca come una specializzazione professionale a sé stante, o come una *skill* integrativa del bibliotecario stesso. In secondo luogo, occorre contestualmente valutare l'adeguatezza o meno dei curricula e dei percorsi didattici attuali. Dalla prima alla seconda considerazione: il *data curator* è un professionista 'ulteriore' di nuovissima generazione, che, in quanto a competenze, si pone al punto di confluenza fra archivistica, biblioteconomia, scienze dell'informazione e informatica. Il suo bisogno formativo richiede pertanto un curriculum apposito di natura marcatamente interdisciplinare5, in seno al quale la sfida maggiore, nonché il più arduo scoglio culturale, è l'abbattimento della frontiera fra il presunto appannaggio umanistico delle discipline LIS e la dimensione delle tecnologie informatiche. La *digital curation*, com'è in effetti ovvio, non si dà senza competenze informatiche; ma se è vero che il *data curator* non si appiattisce sulla figura del bibliotecario, è altrettanto vero che esso non coincide *tout-court* nemmeno con l'ingegnere informatico.

Su questo presupposto si fonda il progetto biennale

[7] Presentazione dal titolo "ACTRIS – Aerosol, Clouds and Trace gases Research Infrastructure"
[8] Presentazione dal titolo "Cultural Heritage: when data are much worst than one can believe"
[9] Presentazione dal titolo "INSTRUCT – Integrated Structural Biology Infrastructure"

europeo EDISON (2015-2017). Andrea Manieri[10] ne illustra il proposito principale: indurre un incremento dell'offerta formativa in *Data Science*, ciò che si realizza attraverso 1) la sistemazione dei precisi contorni di tale professionalità emergente (DSP – Data Science Professional profiles definition), 2) l'individuazione del dominio di conoscenze e competenze di essa identificative (DS-Bok – Data Science Body of Knowledge), e 3) l'adeguamento dei curricula formativi alle richieste specifiche dei settori di impiego interessati (MC-DS – Data Science Model Curriculum). L'Università di Perugia – prosegue Manieri – offre, a questo proposito, il Master di II livello in Data Science, che si propone di formare una figura esperta nella definizione di strategie di business, e facente uso di competenze di ingegneria, informatica, matematica, statistica, economia, gestione aziendale, comunicazione e marketing. Il *data scientist* – conferma Valerio Grossi, che presenta l'offerta formativa ulteriore del Master SoBigData dell'Università di Pisa[11] – è colui che dovrà esser capace di acquisire, analizzare e visualizzare i cosiddetti *big data* (cioè l'ingente quantitativo di dati grezzi inconsapevolmente disseminati in rete da ciascuno di noi, a velocità e in varietà abnormi), di modo da estrarne conoscenza a supporto di decisioni, di iniziative imprenditoriali, sociali, o di analisi predittive. Senza mai dimenticare, beninteso, il vincolo etico e giuridico della priorità della tutela dell'individuo sull'utilità dei dati raccolti.

Sempre a proposito del Master SoBigData, si sofferma ancora, nella prima relazione della seconda giornata di studi, Fosca Giannotti[12], la quale sottolinea che la creazione di un ambiente idoneo per trattare dati sociali debba avvenire integrando infrastrutture di ricerca nazionali, operando attraverso diversi settori e creando una piattaforma interoperabile che permetta l'accesso ai dati da parte di diverse comunità rispettando, al contempo, i vincoli di privacy imposti dal proprietario dei dati. L'estrazione di dati sociali utili, infatti, può avvenire solo in un ambiente che permetta la condivisione dei dati e il loro riutilizzo. A ciò assolvono i VRE (*Virtual Research Environments*), che permettono un'acquisizione dei dati flessibile, condivisibile e sicura, e vengono incontro ai determinati e diversificati bisogni di comunità differenti.

Rendere accessibili i dati sociali apporta senza dubbio un valore aggiunto alla democrazia. Tuttavia si pone in luce un'altra istanza democratica fondamentale: la necessità e il dovere di proteggere gli individui. Viene perciò proposto un metodo di raccolta, analisi e diffusione che tenga conto di principi etici.
Leonardo Sacconi, nel suo intervento *Whole brain optical imaging* si pone invece la questione di come Lens (European Laboratory for Non-Linear Spectroscopy) affronta lo studio di strutture complesse, come le reti neurali, implementando l'uso di diverse tecniche di *imaging*, le quali, offrendo diversi punti di vista, riescono a fornire informazioni complementari sul ruolo dei componenti neurali. L'utilizzo di tecniche come quella della de-convoluzione dell'immagine permette di ottenere, tramite semplici algoritmi di localizzazione a basso costo, ottime prestazioni creando un'immagine più uniforme nella quale le strutture significative sono ben visibili (si parla infatti di de-convoluzione semantica). Risulta così indispensabile l'uso di un DMP che prevede *in primis* il mantenimento di un software fatto su misura (*custom–made software*) idoneo a lavorare con *set* di dati molto grandi (produzioni di dati per settimana di circa cinque terabyte).
Il Progetto RITMARE affrontato da Paola Carrara[13], mette in evidenza la sfida della gestione, distribu-

[10] Presentazione dal titolo "Towards a Community-driven Data Science Body of Knowledge – Data Management Skills and Competences"
[11] Presentazione dal titolo "*Educating Data Scientists: the SoBigData master experience"*

[12] Presentazione dal titolo "*SoBigData. European Research Infrastructure for Big Data and Social Mining"*
[13] Presentazione dal titolo "*Facing data sharing in a heterogeneous research community: lights and shadows in the RITMARE project,"*

zione, pubblicazione dei dati affrontati dal Sottoprogetto 7 di RITMARE, progetto di punta per la ricerca marina nato nel 2012: il progetto fin da subito si preoccupa di realizzare un'interoperabilità che non sia solo tecnologica, bensì anche semantica e sintattica. Seguendo lo schema FAIR, attraverso il suo Sottoprogetto 7, RITMARE assicura e facilita la modifica dei meta-dati, con collegamenti per vocabolari al portale RITMARE (in costruzione); promuove, al fine di favorire lo scambio di dati e l'accesso aperto, la sua *Data Policy*. La *Data Policy* di RITMARE impone l'uso di regole per l'utilizzazione dei dati o prodotti sia agli utenti della comunità sia agli utenti esterni ad essa: 1) la citazione obbligatoria del proprietario/generatore dei dati, 2) l'interrogazione al proprietario/generatore dei dati riguardo alla volontà di partecipare come co-autore entro due anni dalla pubblicazione dei dati, 3) la creazione di un Moratorium (ovvero di un periodo di tempo nel quale i dati vengono resi disponibili e il loro utilizzo è soggetto alle decisioni del proprietario/generatore), 4) i dati grezzi devono essere accompagnati da tutti i dati ancillari (ad es. file di calibrazione) e infine 5) per i dati/prodotti di Background (dati/prodotti resi disponibili dai partner) si applicano le licenze o le regole d'uso a loro associati. RITMARE inoltre promuove licenze *open* sempre nel contesto dell'accessibilità. Riguardo l'interoperabilità, obiettivo principale del progetto RITMARE, è stato realizzato GET-IT Starter Kit (*suite* software originale, fruibile in formato *open*) abilitante i ricercatori alla produzione di propri servizi standard OGC (Open Geospatial Consortium) per la propagazione interoperabile di dati sia osservativi sia geografici (e conseguentemente dei relativi meta-dati) in infrastrutture di dati spaziali. Il riuso dei dati standard OGC è supportato da una comunità mondiale; si sta cercando di elaborare inoltre, per permettere un più facile riutilizzo dei dati, la creazione di identificativi permanenti.

In che modo i meta-dati contribuiscono alla realizzazione di una gestione dati di tipo FAIR? Al quesito tenta di rispondere Eva Méndez, dando anzitutto la definizione di meta-dato: «un dato che fornisce informazioni su una risorsa»[14]. Al di là di questa essenziale formula definitoria, occorre scendere nel dettaglio della relazione che il meta-dato intrattiene con il dato che esso individua e descrive, e chiarirne la specificità. Il meta-dato, come si è detto, è ciò che descrive una risorsa, pertanto esso è un dato a tutti gli effetti, sennonché, rispetto al dato, esso esce da un regime di autoreferenzialità, e si rivolge preminentemente all'utenza, ai servizi software e alle risorse computazionali. Il meta-dato, inoltre, non si limita ad essere utilizzato nel momento della scoperta e della descrizione del dato in questione, ma agisce anche al livello della sua contestualizzazione (lo dimostra, ad es., il caso della rilevanza variabile che un dato può assumere a seconda dei determinati e differenziati contesti di ricerca). Requisito fondamentale del meta-dato, infine, è che esso sia interpretabile sia dalla macchina che dall'uomo.

La questione dell'interoperabilità dei dati e dell'accessibilità e della funzionalità dei meta-dati è certo fondamentale ai fini della gestione dati FAIR. Ma le effettive possibilità di applicazione pratica dei principi FAIR risentono anche di condizionamenti di tipo ulteriore, quali, ad es., le disponibilità di tipo logistico. Giovanni L'Abate, a questo proposito, porta all'attenzione la problematica dell'impossibilità di una preservazione a lungo termine della totalità dei dati utilizzati nel corso di una ricerca[15]. Le ragioni principali di questa impossibilità sono due: 1) l'ingente spazio fisico necessario per immagazzinare i dati; 2) il costo della gestione di questi spazi. Risulta quindi necessario operare una scelta dei dati da conservare e renderli comunque sempre accessibili.

Andrea Ferretti illustra nel suo intervento *Handling data and workflows in computational materials science: the AiiDA initiative* l'iniziativa AiiDA, la

[14] Presentazione dal titolo *"Cool" metadata for FAIR data*".

[15] Presentazione dal titolo "*Soil Research Data Policies, Data availability and Access, and Interoperability challenge for CREA Soil Open Data, Italy"*

quale permette l'avanzamento dell'"High-throughput computing" (HTC), una metodologia efficace nella scienza dei materiali usando metodi computazionali per la scoperta di nuovi materiali, che si sta diffondendo rapidamente e che sta diventando uno strumento essenziale. AiiDA memorizza automaticamente i calcoli con i loro ingressi e le uscite in un *repository* idoneo per garantire la riproducibilità di ogni calcolo, la persistenza dei dati stessi e la rapida interrogazione dei risultati tramite *database* su misura. Lo strumento permette inoltre la codifica, l'esecuzione e la riproduzione non solo dei calcoli singoli bensì anche dei flussi di dati.

Viene affrontato, nella nona sessione del convegno, il problema della citazione dei dati, del controllo delle loro versioni e della loro provenienza. Quali sono le esperienze e le soluzioni attuali?

Seguendo le «Raccomandazioni del gruppo di lavoro RDA sulla citazione dati» (si veda la presentazione *Enabling Precise Identification and Citability of Dynamic Data: Recommendations of the RDA Working Group on Data Citation* di Andreas Rauber) si deve riuscire a identificare con precisione il sottoinsieme di dati utilizzati per poterli applicare da uno studio precedente ad un nuovo modello.

Mentre le descrizioni verbali di come è stato creato il sottoinsieme (ad esempio fornendo la gamma degli attributi selezionati e gli intervalli di tempo) sono raramente precise e non supportano il trattamento automatico, mantenendo copie sovrabbondanti dei dati in questione e non incrementando la regolazione dei *big data*. L'assegnazione di identificatori persistenti ai *set* di dati interi o singoli sottogruppi o elementi di dati, non sono sufficienti a soddisfare queste esigenze. Questo problema è ulteriormente aggravato se nuovi dati continuano ad essere aggiunti a un *database*, dalla correzione o eliminazione di dati. Le soluzioni a queste problematiche sono cercate con approcci non convenzionali basati su fonti di dati identificati con precise coordinate temporali (*time-stamped*) e con modelli, con identificativi permanenti assegnati alle *queries*/istruzioni *time-stamped*, le quali sono utilizzate per la creazione di sottoinsiemi di dati. *Database* in stile SQL, *file* XML e CSV possono a loro volta incrementare queste soluzioni atipiche nella gestione delle problematiche della citazione dei dati.

Quante risorse vengono dissipate nella ricerca? Circa l'85%. Questo è il primo dato sconvolgente emerso *da CoBRA guideline : a tool to facilitate sharing, reuse, and reproducibility of bioresource-based research*. Elena Bravo evidenzia come le cause di questo spreco sono da rintracciarsi nella mancanza di standards per la citazione delle biorisorse (campioni biologici con rispettivi dati) che rende impossibile recuperarle nelle banche dati bibliografiche; questa mancanza è anche la causa principale per cui la competenza dei ricercatori che hanno realizzato le biorisorse risulta ad oggi non riconosciuta.

La necessità di creare questo standard per la citazione delle risorse biologiche per rendere possibile il recupero di articoli o di riviste basate sull'utilizzo di campioni biologici e la loro analisi in letteratura scientifica ha dato vita nel 2015 al progetto CoBRA (Citation of BioResources in journal Articles): un insieme di linee guida per la citazione di risorse biologiche per la letteratura scientifica. L'utilizzo di queste linee guida ha permesso il miglioramento della qualità delle relazioni scientifiche, con riduzione dei costi di ricerca e rispettando il criterio di riproducibilità proprio della scienza. Infine anche il lavoro del ricercatore risulta non solo più riconosciuto ma anche più diffuso.

La descrizione, la citazione, il riuso dei dati di ricerca è fondamentale per rendere la scienza riproducibile; questo è vero in tutti i contesti scientifici ed in particolare in Information Retrieval (la quale si occupa della rappresentazione, memorizzazione e organizzazione dell'informazione testuale). In *Reproducibility for IR evaluation*, Gianmaria Silvello mostra diversi sistemi che si occupano non solo della gestione dei dati IR (come *Direct*) ma anche dell'implementazione della loro visibilità sulla rete (come il progetto LOD – DIRECT).

La riproducibilità è intrinsecamente legata alla citazione dei dati: certo è fondamentale identificare in modo univoco i dati, ma è altresì necessario: 1) predisporre un sistema uomo-macchina che renda intellegibili i frammenti di citazioni; 2) definire uno strumento per creare facilmente le citazioni dei dati; e 3) infine sviluppare un sistema di citazione che includa il più basso coinvolgimento possibile di *data creators* e *curators*.

La decima e ultima sessione del convegno si è occupata infine delle modalità italiane per implementare i DMP nelle varie discipline: D4SCIENCE, presentata in *D4Science Data infrastructure: a facilitator for a FAIR data management* a cura di Pasquale Pagano, risulta essere un'infrastruttura per dati ibridi, ovvero provenienti da diversi ambiti e discipline che spaziano da quelle umanistiche a quelle scientifiche. Le tecnologie in D4Science sono integrate e forniscono un accesso flessibile all'uso delle funzionalità sui dati e alle modalità di implementare il data–management. L' approccio infrastrutturale permette l'abbattimento dei costi mantenendo al contempo la possibilità per i ricercatori una vera e propria scienza open giorno dopo giorno. Nell'ottica FAIR, D4SCIENCE assicura: 1) la reperibilità tramite identificatori unici e usando set di meta-dati, 2) l'accessibilità, condividendo e pubblicando risorse disponibili attraverso protocolli multipli, nonché facilitando 3) l'interoperabilità e arricchendo automaticamente le risorse con meta-dati in formati multipli e infine 4) la ri-usabilità è ottenuta attraverso diversi meccanismi che includono tra l'altro la generazione sistematica della provenienza dei meta-dati.

Con lo sviluppo di *set* di dati e meta-dati, la gestione scientifica degli stessi sta diventando sempre più complessa, non solo per una questione meramente legata all'analisi e all'immagazzinamento degli stessi, quanto anche al problema dell'accesso che deve tenere conto dei diritti dei proprietari e dell'aspetto *open*. Attraverso EuHIT e l'infrastruttura EUDAT si può osservare come si cerchi non sono di produrre, preservare, condividere e riusare dati prodotti da diverse organizzazioni ma che questi dati sono concentrati su una vasta area geografica che implica la necessità di tutelare i diritti dei proprietari al di là dei confini nazionali. Come sottolinea Claudio Cacciari in *Data management experiences in the European projects context: which lessons for us,* è importante che le varie comunità scientifiche vertano verso la ricerca di uno standard condiviso poiché i *data center* che offrono servizi informatici non possono colmare questa lacuna, bensì i *data center* devono offrire servizi e applicare politiche che sostengono i principi FAIR in modo flessibile.

Il Consortium GARR si occupa di progettare, implementare e di far operare una infrastruttura di rete che permetta al mondo accademico e scientifico di utilizzare gli strumenti comunicativi al fine di realizzare le loro attività di ricerca e insegnamento in ambito nazionale e non. Oltre a questo, esso ha recentemente sviluppato un *Cloud Federated*, presentato da Giuseppe Attardi in *A Reference Architecture for a Federated Cloud for Research*, il quale si pone come obiettivi la condivisione di risorse preservando la proprietà in uno spazio dove diverse organizzazioni di ricerca possano contribuire con porzioni dei loro *cloud* per creare una cooperazione basata sulla condivisione e il rispetto della proprietà intellettuale.

Come si può rendere infine efficace e efficiente un DMP? Questi sono gli interrogativi finali ai quali ha cercato di rispondere Davide Salomoni presentando *Efficient and effective: can we combine both to realize high-value, open, scalable, multi-disciplinary data and compute infrastructures?* Indigo Data Cloud si pone infatti i seguenti obiettivi: 1) lo sviluppo di soluzioni aperte e interoperabili per i dati scientifici; 2) il supporto alla scienza *open* creando uno spazio dati europeo e infine 3) rendere possibile la collaborazione di diverse comunità scientifiche di ogni parte del mondo. Indigo ottiene risultati concreti allineando la visione delle diverse comunità di ricerca con le raccomandazioni correnti, in special modo quelle di RDA. Pur rimanendo molto complesso combinare efficienti e effettive soluzioni

quando si cerca di sfruttare dati distribuiti/risorse computazionali, Indigo si sta focalizzando sul consolidamento del suo software, aggiungendo nuove caratteristiche e distribuendolo nelle e-infrastrutture.

Le due giornate si concludono, dunque, con un focus sullo stato dell'arte in Italia in ambito di strategie e proposte di infrastrutture destinate alla redazione implementata di DMP.

Ma più che delle conclusioni definitive, dal workshop emerge una questione aperta e una sfida concreta: quali sono i bisogni specifici da soddisfare, qual è il piano di azione a medio termine più efficace per un intervento graduale e specifico su ciascuna delle lacune che l'Italia al proposito presenta? La domanda è aperta. O meglio è open: accessibile e ad uso dell'ingegno di chiunque voglia, possa e debba offrire un contribuito fattivo alla creazione, alla conservazione e alla diffusione della conoscenza.