Paolo Manghi

Sfide tecnologiche per l'accesso aperto a tutti i prodotti della ricerca

Abstract

Scientific communication workflows, traditionally conceived to serve publishing of literature, are today evolving in order to meet novel requirements of science. Indeed, scientists are increasingly interested to share and discover the narration of research results together with the actual research datasets and experiments described in the article. Facing Open Access challenges for all products of science means overcoming technological and access rights not only for scientific publications, but also for the datasets and the experiments associated to them.

1. L'evoluzione della comunicazione scientifica

Con il termine "infrastrutture per la ricerca" s'intende l'intero complesso degli elementi organizzativi (ruoli, procedure, ecc.), strutturali (edifici, laboratori, ecc.) e tecnologici (microscopi, telescopi, sensori, risorse ICT, ecc.) a supporto dei processi investigativi e sperimentali degli scienziati. La parte ICT di queste infrastrutture è divenuta oggigiorno essenziale per fruire dei flussi di comunicazione scientifica. Da un lato gli scienziati depositano i file e i metadati delle pubblicazioni in "publication repository" istituzionali o tematici o presso riviste online ("journal"). In seguito, vari applicativi web svolgono l'utile compito di aggregare, indicizzare e catalogare queste sorgenti di informazione, allo scopo di offrire ai ricercatori servizi avanzati per la ricerca della letteratura (es. Google Scholar, DBPL).

Tuttavia, rispetto al passato, la ricerca moderna pone un forte accento sull'elaborazione di grandi quantità di dati (Lynch, 2009), portando gli scienziati a investire fondi ed energie per collezionare, curare e produrre dati per la ricerca.

Nell'ultimo decennio, questo trend ha spinto i dati a ricoprire il ruolo di "cittadini di prima classe" nell'ambito della comunicazione scientifica, alla stessa stregua della letteratura. A supporto e riprova di questo processo evolutivo si è assistito alla diffusione di "data repository" (GigaScience, Dryad, FigShare, Pangaea, ecc.), di varie iniziative di standardizzazione per la citazione dei dati (DataCite, Dataverse, Force11, ecc), e, in senso più ampio, alla definizione di pratiche scientificamente riconosciute per la pubblicazione di dati sperimentali.

A rafforzare e completare questa rivoluzione di paradigma, un'altra sfida tecnologica ha di recente suscitato vivo interesse: la pubblicazione degli esperimenti (De Roure et al., 2010), intesi come i processi metodologici o i flussi ("workflow") necessari a trarre determinate conclusioni scientifiche. L'obiettivo è di fornire ai ricercatori gli strumenti necessari per ripetere ("stesso esperimento, stesso laboratorio"), replicare ("stesso esperimento, altro laboratorio"), riprodurre ("stesso esperimento, diversa configurazione") o riusare ("includere parte dell'esperimento in altro esperimento") l'esperimento, massimizzando quindi la trasparenza e il riuso dei risultati scientifici.

Introdurre nei flussi di comunicazione scientifica la possibilità di pubblicare letteratura unitamente a dati ed esperimenti a essa correlati consentirebbe:

1. una miglior interpretazione dei risultati scientifici;
2. l'applicazione di più rigorosi criteri di valutazione del lavoro scientifico, riducendo la possibilità di "frodi";
3. l'introduzione di criteri di premiazione omnicomprensivi del merito scientifico;
4. la riduzione dei costi della ricerca, promuovendo il riuso.

2. L'accesso aperto a tutti i prodotti della ricerca

Gli innegabili vantaggi dell'accesso aperto alla letteratura scientifica sono noti ai più e possono essere brevemente elencati come: pari opportunità nello svolgimento della ricerca, riduzione dei costi alla ricerca pubblica (Houghton et al., 2009), aumento della produzione scientifica (Willinsky, 2005), aumento delle citazioni (Wagner, 2010; Opcit Project, 2012 [1]) e coinvolgimento del pubblico non-scientifico (Swan, 2010).

In virtù dei sopra descritti nuovi requisiti della comunicazione scientifica, la comunità ha quindi iniziato a porsi il problema di come applicare lo stesso paradigma ad altre tipologie di prodotti per la ricerca, quali i dati e gli esperimenti sopra citati. Ad esempio, il programma di finanziamento Horizon2020 della Commissione Europea ha introdotto il "Data Pilot", un'attività che ha come obiettivo la regolazione degli obblighi riguardanti la pubblicazione, la persistenza e il libero accesso ai dati della ricerca prodotti nel contesto di progetti finanziati. Tuttavia, dopo un'attenta analisi, il problema risulta essere diverso da quello noto per la letteratura, e la sfida decisamente più ardua. La definizione fornita dalla Commissione per l'accesso aperto cita:

Open Access can be defined as the practice of providing on-line access to scientific information that is free of charge to the end-user and that is re-usable [European Commission].

Da essa si evince che per realizzare l'accesso aperto a un prodotto della ricerca è necessario garantire due proprietà: "*…online access […] and that is re-usable*" e "*…is free of charge to the end-user…*". La prima proprietà è di carattere tecnologico, ed è garantita dal consolidamento di rigorosi e riconosciuti (in certi casi certificati) flussi di comunicazione scientifica per i prodotti: la sequenza delle quattro fasi di "submission" (tipicamente on-line), "peer-review" (tipicamente "single/double blind", ma di recente anche "self-archiving"), "ricerca" e "citazione".

La seconda proprietà è invece negoziata a livello di politiche e diritti d'accesso, ha quindi un forte carattere legale e giuridico. Tendenzialmente non ci si pone il problema di garantirla per prodotti che non soddisfano la prima proprietà. Ad esempio, come introdotto nella Sezione 1, per la letteratura scientifica i flussi di comunicazione scientifica sono supportati da tecnologie e pratiche consolidate. Le iniziative per l'Accesso Aperto si occupano quindi della risoluzione delle problematiche inerenti i diritti d'accesso. Al contrario, la maturità dei flussi di comunicazione scientifica per dati ed esperimenti è ancora ad un livello embrionale se comparata a quella della letteratura.

In molte comunità scientifiche le pratiche di "submission" on-line dei dati o degli esperimenti sono vaghe o totalmente assenti, quindi tali prodotti rimangono a tempo indeterminato negli hard-disk, negli appunti, o nelle menti dei ricercatori. Di conseguenza, le sfide ad attuare l'accesso aperto a questi prodotti non sono da imputare unicamente a problemi di carattere legislativo o economico, ma anche a problematiche di ordine culturale e tecnologico.

3. Flussi di comunicazione scientifica per dati ed esperimenti

Come sopra accennato, per la letteratura scientifica i flussi sono ben stabiliti, con qualche idiosincrasia tipica delle singole discipline. I ricercatori inviano un articolo a una conferenza o a una rivista scientifica, l'articolo è revisionato secondo pratiche condivise (single-blind, double-blind, self-archiving, ec.), se accettato, l'articolo viene depositato in formato elettronico nel repository istituzionale o tematico di riferimento e al sito della rivista scientifica, se disponibile. La ricerca on-line della letteratura è effettuata tramite aggregatori di carattere più o meno tematico, come Google Scholar, OpenAIRE, Scopus, ecc. Le pratiche di citazione della letteratura sono infine ben stabilite grazie a standard di metadati (Dublin Core, Bibtex, MARC, ecc.) che tendenzialmente prescindono dalla disciplina. Ben diversa è la situazione riguardante dati ed esperimenti.

Dati della ricerca. In assenza di adeguati meccanismi d'attribuzione del merito, in molte discipline la condivisione e pubblicazione dei dati arrivano a essere ritenute controproducenti. I flussi di comunicazione scientifica sono invece ben stabiliti nell'ambito delle discipline scientifiche in cui il riuso dei dati è ritenuto storicamente importante (es. biodiversità) o in quelle in cui la coesione scientifica e le infrastrutture per la ricerca hanno permesso il cambiamento (es. INSPIRE, dati geo-spaziali). In virtù del carattere disciplinare dei dati e delle loro molteplici forme di riuso, è complesso arrivare a condividere flussi di carattere inter-disciplinare come nel caso della letteratura. Molte iniziative stanno oggi studiando il problema (Force11, RDA, ecc.) e le soluzioni a peer-review (es. assente, manuale, semi-automatica) e deposito digitale (es. "data repository" in appositi "data centres") sono le più svariate, così come quelle offerte per la ricerca on-line (es. DataCite e "data repository" tematici) e gli standard per la citazione (Data Cite, Dataverse, CERIF, ecc.).

Esperimenti. Se per i dati il problema della comunicazione è spesso contestualizzato a una data comunità o disciplina scientifica, per gli esperimenti il problema può avere la grana ancor più fine del laboratorio o dell'infrastruttura di ricerca. Data la complessità della questione, non esistono a oggi flussi di comunicazione scientifica per esperimenti che siano accreditati da una comunità scientifica. Di conseguenza, tutte le attività a contorno sono di carattere investigativo o sperimentale. Per la peer-review e il deposito di esperimenti in formato digitale riferimenti ormai noti sono il lavoro fatto da "myexperiments.org" (research objects, wf4ever), le attività svolte nel contesto degli "executable papers" e quelle relative agli appunti sperimentali tipo E-notebook.

Le comunità scientifiche sono oggi impegnate a definire, consolidare e stabilire precisi flussi di comunicazione scientifica per dati ed esperimenti, allo scopo di identificare le pratiche e le tecnologie per la valutazione della qualità (peer-review), il deposito e la ricerca on-line, e la citazione. Nel seguito introduciamo alcune tra le iniziative più note e interessanti in questi settori.

3.1. Contestualizzare prodotti di ricerca all'articolo tradizionale

L'approccio è quello di sfruttare i flussi di comunicazione scientifica per la letteratura introducendo però nuove forme "dedicate" di articolo scientifico, orientate a descrivere prodotti specifici (dati, esperimenti, software) o a "incorporare" alla narrazione altri tipi di prodotti della ricerca mediante

tecniche Web 2.0.

Riviste dedicate: le soluzioni adottate in questo settore hanno come idea di fondo quella di pubblicare dati ed esperimenti in modo indiretto, sfruttando i flussi di comunicazione scientifica per la letteratura. Ad esempio, per i dati si assiste a due pratiche, con relative tecnologie a supporto:

- L'obbligo di depositare on-line i dati in un "data repository" per poter sottoporre un articolo a peer-review e, in caso di accettazione dell'articolo, l'obbligo a mantenere un riferimento dalla pubblicazione ai dati e viceversa (es. Joint Data Archiving Policy, JDAP);
- La nascita delle riviste di tipo "data journals" dedicate alla pubblicazione dei dati, nelle quali l'articolo descrive aspetti riguardanti la creazione dei dati della ricerca e contiene un riferimento ai dati depositati e accessibili on-line.

Enhanced publication: altra categoria di soluzioni è quella che riguarda le cosiddette "enhanced publication" (Bardi, 2014), più genericamente "compound objects". Queste sono pubblicazioni digitali che comprendono, oltre alla parte narrativa dell'articolo tradizionale, anche i dati scientifici, gli esperimenti, il software, ogni tipo di prodotto ad essa correlato. In alcuni casi i prodotti sono distribuiti unitamente alla pubblicazione, come una sorta di pacchetto informativo, in altri casi si sfruttano riferimenti a oggetti remoti. Le "enhanced publication" sono poi soggette a flussi di comunicazione tipici della letteratura, sfruttando quindi meccanismi oliati e riconosciuti per la pubblicazione.

In entrambi i casi sopra descritti, il problema rimane comunque quello della peer-review e del riuso. Valutare la qualità di prodotti come dati ed esperimenti non è un'attività che un ricercatore può sempre eseguire con successo e completezza, in mancanza di contesto e strumenti. Queste inconsistenze sono dovute all'uso improprio dei flussi di comunicazione scientifica per la letteratura, concepiti per la valutazione e il riuso della narrazione scientifica. La seguente categoria di soluzioni tende a colmare queste problematiche e ad affrontare il problema più radicalmente.

3.2 Nuovi flussi di comunicazione scientifica

Un approccio alternativo è quello che mira ad abbandonare i modelli tradizionali, al fine di servire una scienza che è cambiata nella forma e nella sostanza. La motivazione è di fondere il luogo in cui la ricerca si svolge e quello in cui la ricerca si pubblica, fornendo tecnologie che anticipino, semplifichino e quindi accelerino i processi di pubblicazione. Di seguito descriviamo due soluzioni diametralmente opposte: la prima ha come obiettivo quello di immergere i flussi di comunicazione scientifica all'interno delle infrastrutture della ricerca, la seconda ha come scopo quello di realizzare i flussi di comunicazione scientifica unificando servizi e prodotti offerti da diverse infrastrutture per la ricerca.

Science 2.0 Repository. Questa categoria di repository (Assante et al., 2015) sposa il principio che la ricerca è un'attività in corso d'opera che produce un insieme di prodotti tutti utili ai fini della comprensione, riuso e valutazione della ricerca stessa. Pertanto un repository non può inteso solamente come un luogo dove vengono "congelate" descrizioni narrative di attività a un certo istante nel tempo (articolo tradizionale in un repository istituzionale) o dataset completamente decontestualizzati dal contesto dell'esperimento (dati in data repository), ma un luogo dove è possibile accedere all'intero storico dei prodotti della ricerca, in tutte le loro versioni e interconnesioni semantiche tra di loro, scambiandosi informazioni in una logica Web 2.0.

A tal fine il repository deve essere integrato alle infrastrutture per la ricerca, in modo da

"intercettare" automaticamente la generazione di prodotti (senza necessariamente lasciare questo compito ai ricercatori), tenendo traccia dell'ambiente in cui sono stati generati, e in seguito mediarne lo scambio con il mondo. Un esempio di Science 2.0 Repository è stato realizzato nell'infrastruttura per la ricerca iMarine ed è oggi utilizzato dai ricercatori della FAO (Assante et al., 2014) per lo scambio di dati, esperimenti e feedback.

Realizzazione di infrastrutture per la comunicazione scientifica. Le soluzioni in quest'ambito partono dal principio che le infrastrutture per la ricerca hanno pratiche d'uso per la disseminazione scientifica consolidate e non sono generalmente predisposte al cambiamento, per questioni culturali ma anche e soprattutto economiche (la disseminazione scientifica non è una responsabilità e un costo che viene attribuito alle infrastrutture).

Le infrastrutture per la comunicazione scientifica (Castelli, 2013) sono quindi sistemi per la realizzazione di flussi di comunicazione scientifica ottenuti come "ecosistemi di infrastrutture della ricerca". Offrono servizi per l'integrazione di strumenti e contenuti generati in tali infrastrutture sfruttando gli elementi di interoperabilità già offerti da quest'ultime, e permettono la costruzione di flussi di comunicazione scientifica alternativi (ad esempio il riuso di servizi e strumenti di una infrastruttura per permettere la peer-review di dati scientifici). Le infrastrutture OpenAIRE per l'Europa e SHARE per gli Stati Uniti sono esempi di queste infrastrutture.

4. Conclusioni

Lo scopo di questo lavoro è mostrare come le sfide alla realizzazione dell'accesso aperto a tutti i prodotti della ricerca, quindi dell'Open Science, non siano esclusivamente di carattere legale e giuridico come in molti tendono a pensare. Accrescere una cultura dell'accesso aperto è anche e soprattutto una questione sociale che deve essere affrontata dagli scienziati, almeno per coloro che ancora non ne colgono i vantaggi, e adeguatamente supportata da tecnologie e pratiche che ancora stentano a essere identificate e/o abbracciate.

Infatti, sono molte le iniziative di carattere fondazionale che tentano di suggerire rappresentazioni digitali comuni per dati sperimentali ed esperimenti (Research Objects exchange format); di definire i relativi formati di metadati per la citazione e il riuso; di stabilire politiche per l'accesso aperto e la condivisione (es. Joint Declaration of Data Citation Principles, Force11 [2]) (FAIR data principles, Force11 [3]). Inoltre, i flussi di comunicazione scientifica devono essere necessariamente rivisti in modo da adeguarsi ai requisiti della scienza moderna. Questo passaggio, necessariamente tecnologico e culturale prima che giuridico, apre scenari interessantissimi per la ricerca nel settore dell'informazione digitale e delle infrastrutture per la scienza.

Paolo Manghi, Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" - Consiglio nazionale delle Ricerche, Pisa, e-mail: paolo.manghi@isti.cnr.it

Bibliografia

(Assante et al., 2014): Massimiliano Assante - Leonardo Candela - Donatella Castelli - Francesco Mangiacrapa - Pasquale Pagano, *A Social Networking Research Environment for Scientific Data Sharing: The D4Science Offering*, "The Grey Journal", 10.2 (2014), p. 65-71.

(Assante et al., 2015): Massimiliano Assante - Leonardo Candela - Donatella Castelli - Paolo Manghi - Pasquale Pagano, *Science 2.0 Repositories: Time for a Change in Scholarly Communication*, accepted for publication on "D-Lib Magazine", Special Issue January 2015.

(Bardi, 2014): Alessia Bardi - Paolo Manghi, *Enhanced Publications: Data Models and Information Systems*, "Liber Quarterly" 22 (2014), <https://liber.library.uu.nl/index.php/lq/article/view/8445>.

(De Roure et al., 2010): David De Roure et al., *Towards open science: the myExperiment approach*, "Concurrency and Computation: Practice and Experience", 22.17 (2010), p. 2335-2353.

(Castelli, 2013): Donatella Castelli - Paolo Manghi - Costantino Thanos, *A vision towards Scientific Communication Infrastructures*, "International Journal on Digital Libraries", 13.3-4 (2013), p. 155-169.

(Candela et al, 2015): L. Candela - D. Castelli - P. Manghi - A. Tani, *Data Journals: a Survey*, "Journal of the Association for Information and Science Technology", accepted for publication in June 2014. Will appear with DOI: 10.1002/asi.23358.

(Houghton et al., 2009): John Houghton - Peter Sheehan, *Estimating the potential impacts of open access to research findings*, "Economic Analysis and Policy", 39.1 (2009), p. 127-142.

(Lynch, 2009): C. Lynch, *Jim Gray's fourth paradigm and the construction of the scientific record*, in T. Hey - S. Tansley - C. Tolle (eds.), *The Fourth Paradigm*, Redmond, Microsoft Corporation, 2009, p. 177–183.

(Swan, 2010): Alma Swan, *The Open Access citation advantage: Studies and results to date*, 2010, <http://eprints.soton.ac.uk/268516/>.

(Wagner, 2010): Benno Wagner, *Open access citation advantage: an annotated bibliography*, "Issues in Science and Technology Librarianship", 60 (2010), 2.

(Willinsky, 2005): John Willinsky, *Scholarly associations and the economic viability of open access publishing*, "Open Journal Systems Demonstration Journal", 1.1, (2005).

Note

[1] Opcit project, <http://opcit.eprints.org/oacitation-biblio.html>

[2] Data Citation Principles, <http://www.force11.org/datacitation>.

[3] FAIR data principles, <https://www.force11.org/group/fairgroup/fairprinciples>.